

DOMANDE FAC-SIMILE ESAME

Qui di seguito, sono riportate alcune delle potenziali domande d'esame. Non verranno pubblicate le soluzioni di ciascuna di essa in quanto alcune di esse sono già state proposte nel materiale a disposizione su Moodle.

CAPITOLO 2 "Richiami di Probabilità"

DOMANDA 1.

Quali sono le possibili fonti di dati? Come si possono poi dividere?

Elencare le diverse tipologie e dare una breve descrizione di ciascuna di essa.

DOMANDA 2.

La tabella seguente fornisce la distribuzione di probabilità congiunta dello stato occupazionale e del livello scolastico per individui impiegati oppure in cerca di occupazione (disoccupati) appartenenti alla popolazione in età lavorativa del Paese A.

	Disoccupato (Y=0)	Occupato (Y=1)	Totale
Non laureato (X=0)	0,078	0,673	0,751
Laureato (X=1)	0,042	0,207	0,249
Totale	0,12	0,88	1,000

- Si calcoli $E(Y)$.
- Il tasso di disoccupazione è la frazione della forza lavoro che è disoccupata. Si mostri che il tasso di disoccupazione è dato da $1 - E(Y)$.
- Si calcoli $E(Y|X = 1)$ e $E(Y|X = 0)$.
- Si calcoli il tasso di disoccupazione per (i) i laureati e (ii) i non laureati.
- Un membro di questa popolazione selezionato a caso dichiara di essere disoccupato. Qual è la probabilità che questo lavoratore sia (i) laureato, (ii) non laureato?
- Livello d'istruzione e stato occupazionale sono indipendenti? Si argomenta la risposta.

DOMANDA 3.

Si illustrino i concetti di distribuzione di probabilità e distribuzione di probabilità cumulata per una variabile casuale discreta X .

Vi sono differenze rispetto al caso di una variabile casuale continua?

Si considerino ora due variabili casuali discrete, X e Y , e si definiscano i concetti di distribuzione di probabilità congiunta, marginale e condizionata.

Si spieghi infine quando X e Y si dicono essere indipendenti.

DOMANDA 4.

Fornire delle definizioni per valore atteso, varianza e deviazione standard.

Definire la distribuzione di probabilità, il valore atteso, la varianza e la deviazione standard nel caso particolare di una variabile casuale di Bernoulli.

DOMANDA 5.

Siano X e Y variabili casuali discrete con la seguente distribuzione congiunta:

		Valore di Y				
		2	4	6	8	10
Valore di X	3	0,04	0,09	0,03	0,12	0,01
	6	0,10	0,06	0,15	0,03	0,02
	9	0,13	0,11	0,04	0,06	0,01

Ovvero, $Pr(X = 3, Y = 2) = 0,04$ e così via.

- Calcolare la distribuzione di probabilità di Y e la distribuzione di probabilità di X .
- Calcolare la distribuzione di probabilità di Y data $X = 3$.
- Calcolare la distribuzione di probabilità di Y data $X = 6$.
- Calcolare la distribuzione di probabilità di Y data $X = 9$.

DOMANDA 6.

Si supponga che le variabili casuali X e Y siano indipendenti e che se ne conosca la distribuzione. Spiegare perchè la conoscenza del valore di X non rivela nulla circa il valore di Y .

DOMANDA 7.

Si usi la distribuzione di probabilità nella Tabella per calcolare:

	Con pioggia ($X = 0$)	Senza pioggia ($X = 1$)	Totale
Percorrenza lunga ($Y = 0$)	0,15	0,07	0,22
Percorrenza breve ($Y = 1$)	0,15	0,63	0,78
Totale	0,30	0,70	1,00

- a) $E(Y)$ e $E(X)$;
- b) σ_X^2 e σ_Y^2 ;
- c) σ_{XY} e $\text{corr}(X, Y)$.

DOMANDA 8.

Si calcolino le seguenti probabilità:

- a. $\Pr(Y \leq 5)$, con Y distribuito come una $N(4, 9)$
- b. $\Pr(Y > 2)$, con Y distribuito come una $N(5, 16)$

DOMANDA 9.

Si stima che il 30% degli adulti negli Stati Uniti siano obesi, che il 3% siano diabetici e che il 2% siano sia obesi che diabetici. Determina la probabilità che un individuo scelto casualmente

- a) sia diabetico se è obeso;
- b) sia obeso se è diabetico.

DOMANDA 10.

Nel censimento del 2000 ogni persona residente negli USA doveva scegliere da un lungo elenco la propria razza. La categoria "Ispanico/latino" è un caso a parte poiché in essa vi possono essere tante razze diverse. Se scegliamo un residente negli USA in modo casuale, in base ai dati del censimento del 2000 abbiamo le seguenti probabilità:

	Ispanici	Non ispanici
Asiatici	0.000	0.036
Neri	0.003	0.121
Bianchi	0.060	0.691
Altro	0.062	0.027

- a) Verifica che questa tabella di probabilità sia corretta.
- b) Quanto vale la probabilità che un americano scelto in modo casuale sia ispanico?
- c) I bianchi di origine non ispanica rappresentano da sempre la maggioranza di residenti negli USA. Quale è la probabilità che un americano scelto in modo casuale non sia membro di questo gruppo?

CAPITOLO 3 “Richiami di Statistica”

DOMANDA 1.

Il dottore di Anna è preoccupato che lei possa soffrire di diabete gestazionale (alto livello di glucosio nel sangue durante la gravidanza). È presente una certa variabilità sia nel reale livello di glucosio nel sangue, sia nei risultati del test che lo misura.

Una paziente è affetta da diabete gestazionale se il livello di glucosio, un'ora dopo aver ingerito una bevanda zuccherata, è superiore ai 140 milligrammi per decilitro (mg/dl). Il livello di glucosio di Anna varia secondo una distribuzione Normale con media $\mu = 125 \text{ mg/dl}$ e $\sigma = 10 \text{ mg/dl}$.

- a) Se si fa una singola misurazione di glucosio, qual è la probabilità che ad Anna sia diagnosticato il diabete gestazionale?

DOMANDA 2.

Qual è la differenza tra ipotesi nulla e ipotesi alternativa?

Tra livello minimo, livello di significatività e potenza?

Tra ipotesi alternativa unilaterale e bilaterale?

DOMANDA 3.

Si consideri il caso di un test d'ipotesi con alternativa bilaterale ad un livello di significatività del 5%.

Quali sono gli steps per condurre un test d'ipotesi circa la media della popolazione?

Come cambierebbe il test d'ipotesi nel caso di un livello di significatività del 10%?

DOMANDA 4.

In una popolazione $\mu_y = 75$ e $\sigma_y^2 = 45$. Si usi il teorema limite centrale per trovare:

$$\Pr(\bar{Y} < 73), \text{ in un campione casuale di numerosità } n = 50$$

DOMANDA 5.

A un campione casuale di 1500 studenti dell'ultimo anno di una scuola superiore viene data una nuova versione del SAT (acronimo dall'inglese Stanford Achievement Test). La media campionaria del punteggio è 1230 e la sua deviazione standard campionaria è 145. Si costruisca un intervallo di confidenza al 95% per la media del punteggio nella popolazione degli studenti dell'ultimo anno della scuola superiore.

DOMANDA 6.

L'econometria fa uso di tre tipologie di metodi statistici: la stima, la verifica di ipotesi e gli intervalli di confidenza. Fornire delle definizioni per ciascuno di essi.

DOMANDA 7.

Nella statistica, gli stimatori possibili sono molti, ma cosa rende uno stimatore "migliore" rispetto ad un altro? Quali sono dunque le proprietà degli stimatori? Si argomenta la risposta.

DOMANDA 8.

Si supponga che un'impresa specializzata nella produzione di lampadine produca lampadine con una vita media di 1000 ore e una deviazione standard di 100 ore. Un inventore dichiara di aver ideato un processo innovativo che permette di produrre lampadine con una vita media più lunga e la stessa deviazione standard. Il manager dell'impresa seleziona a caso 50 prodotti attraverso questa tecnica ed afferma che crederà alle affermazioni dell'inventore se la loro vita media campionaria supererà le 1100 ore, altrimenti concluderà che questo nuovo processo non è migliore di quello vecchio. Sia μ la media del nuovo processo. Si consideri l'ipotesi nulla $H_0: \mu = 1000$ contro l'ipotesi alternativa $H_1: \mu > 1000$.

- Qual è il livello minimo del test usato dal manager dell'impresa?
- Si supponga che il nuovo processo sia effettivamente migliore e che produca lampadine con una vita media di 1150 ore. Qual è la potenza della procedura di test usata dal manager?
- Quale procedura di test dovrebbe usare il manager se volesse ottenere un livello minimo dell'1%?

DOMANDA 9.

Che differenza c'è tra uno stimatore non distorto e uno stimatore consistente?

DOMANDA 10.

Perché si dice che sia fondamentale il campionamento casuale?

DOMANDA 11.

Perché un intervallo di confidenza contiene più informazioni rispetto al risultato di un semplice test d'ipotesi?

DOMANDA 12.

Dalla popolazione dei docenti universitari è stato estratto un campione casuale di 100 docenti di sesso femminile rilevandone l'età.

- a) Determinare l'intervallo di confidenza a livello 95% per l'età media, sapendo che l'età media del campione delle 100 donne osservate è pari a 42.2 e che nella popolazione dei docenti di sesso femminile la variabile età presenta distribuzione Normale con varianza pari a 49;
- b) Si vuole verificare l'ipotesi che l'età media sia pari a 44 anni contro l'ipotesi alternativa bilaterale. Cosa possiamo concludere a livello di significatività 0.05? E se il livello di significatività fosse 0.1?

DOMANDA 13.

Supponiamo che la vita (in ore) di una lampadina da 75 watt sia approssimativamente normalmente distribuita ed abbia una deviazione standard pari a $\sigma = 25$ ore. Un campione di 20 lampadine ha una media (campionaria) di vita $\bar{x} = 1014$ ore.

- a) È ragionevole supporre che la media di vita delle lampadine superi 1000 ore? Usare un livello di significatività $\alpha = 5\%$. E a un livello di significatività dell'1%?

CAPITOLI 4-5 “Regressione Lineare con un Singolo Regressore – Verifica di Ipotesi e Intervalli di Confidenza”

DOMANDA 1.

Elencare le assunzioni dei minimi quadrati nel caso di regressione con un singolo regressore.

Cosa s'intende con omoschedasticità e eteroschedasticità?

Cosa cambierebbe in termini di assunzioni dell'OLS se vi fosse omoschedasticità?

DOMANDA 2.

Un ricercatore conduce un esperimento per misurare l'impatto di un breve sonnello sulla memoria. Si chiede a 200 partecipanti di fare un sonnello di 60 minuti o di 75 minuti. Al risveglio, ciascuno dei partecipanti svolge un test sulla memoria a breve termine. A ogni studente è assegnato casualmente uno dei tempi per l'esame in base al lancio di una moneta.

Sia Y_i il numero dei punti ottenuti nell'esame dall' i -esimo partecipante ($0 \leq Y_i \leq 100$), sia X_i la durata del sonnello fatto dal partecipante prima di sottoporsi al test ($X_i = 60$ o 75) e si consideri il modello di regressione $Y_i = b_0 + b_1X_i + u_i$.

- Si spieghi che cosa rappresenta il termine u_i . Perché diversi partecipanti hanno diversi valori di u_i ?
- Che cos'è $E(u_i | X_i)$? I coefficienti stimati sono non distorti?
- Quali preoccupazioni si dovrebbero avere per quanto riguarda la conformità all'esperimento dei partecipanti?
- La regressione stimata è $\hat{Y}_i = 55 + 0,17X_i$.
 - Si usi la regressione stimata per predire il punteggio medio dei partecipanti che hanno dormito 60 minuti prima di svolgere il test; poi per quelli che hanno dormito 75 minuti e 90 minuti.
 - Si calcoli qual è l'aumento stimato nel punteggio per un partecipante a cui sono dati ulteriori 5 minuti per l'esame.

DOMANDA 3.

Che cos'è una variabile dummy, o una variabile indicatrice?

Si descrivano le differenze nell'interpretazione dei coefficienti di una regressione lineare quando la variabile indipendente è continua e quando è binaria.

Che differenze ci sono nella costruzione di intervalli di confidenza e test di ipotesi quando la variabile indipendente è binaria e quando è continua?

DOMANDA 4.

Si supponga che un ricercatore, utilizzando i dati sulla dimensione delle classi (CS) e il punteggio medio in un test per 50 classi relative al terzo livello d'istruzione, stimi la regressione OLS,

$$\widehat{TestScore} = 640,3 - 4,93 * CS, \quad R^2 = 0,11 \quad SER = 8,7$$

- Una classe è composta da 25 studenti. Qual è la predizione della regressione per quanto riguarda il punteggio medio nei test per tale classe?
- L'anno scorso una certa classe era composta da 21 studenti e quest'anno da 24. Qual è la predizione della regressione per quanto riguarda una variazione nel punteggio medio nei test per la classe?
- Si costruisca un intervallo di confidenza di livello 95% per β_1 , la pendenza della regressione.
- Di calcoli il valore-p di un test bilaterale per l'ipotesi nulla $H_0: \beta_1 = 0$. Si rigetta l'ipotesi nulla (i) al livello di significatività del 5%? (ii) Al livello dell'1%?
- Si calcoli il valore-p per un test bilaterale dell'ipotesi nulla $H_0: \beta_1 = -5.0$. Senza calcoli aggiuntivi, si determini se -5.0 è contenuto nell'intervallo di confidenza al 95% per β_1 .

DOMANDA 5.

Si consideri il caso di un test d'ipotesi con alternativa bilaterale ad un livello di significatività del 5%.

Quali sono gli steps per condurre un test d'ipotesi circa la pendenza β_1 ?

Come cambierebbe il test d'ipotesi nel caso di un livello di significatività dell'1%?

DOMANDA 6.

Una regressione della retribuzione media settimanale (AWE – average weekly earnings, misurata in dollari) sull'età (Age, misurata in anni), usando un campione casuale di lavoratori a tempo pieno laureati compresa di età compresa fra i 25 e i 65 anni, fornisce:

$$\widehat{AWE} = 696,7 + 9,6 * Age, \quad R^2 = 0,023 \quad SER = 624,1$$

- Si spieghi che cosa indicano i valori 696,7 e 9,6 dei coefficienti.
- Cosa rappresentano R^2 e SER ? Quali sono i loro valori?
- Qual è la retribuzione predetta dalla regressione per un lavoratore di 25 anni?
- E per uno di 45?

- e. La regressione dà predizioni affidabili per un lavoratore di 99 anni? Si argomenta la risposta.

DOMANDA 7.

Si supponga che un ricercatore, usando i dati su 200 lavoratori maschi selezionati casualmente e 240 lavoratrici femmine, stimi la regressione OLS

$$\widehat{Wage} = 10,73 + 1,78 * Male, \quad R^2 = 0,09 \quad SER = 3,8$$

(0,16) (0,29)

dove *Wage* è misurato in \$/ora e *Male* è una variabile binaria che è uguale a 1 per un maschio e a 0 per una femmina. Si definisca la differenza salariale di genere come la differenza tra il salario medio di maschi e femmine.

- Qual è la differenza di genere stimata?
- La differenza di genere è significativamente diversa da zero? (Si calcoli il valore-p per verificare l'ipotesi nulla che non ci sia alcuna differenza di genere).
- Si costruisca un intervallo di confidenza al livello 95% per la differenza di genere.
- Nel campione, qual è la retribuzione media delle donne? E degli uomini?
- Un altro ricercatore usa gli stessi dati, ma effettua una regressione di *Wage* su *Female*, una variabile che è uguale a 1 per le femmine e 0 per i maschi. Quali sono le stime ottenute da tale regressione?

DOMANDA 8.

In media, i lavoratori con livello di istruzione più elevato hanno retribuzioni più elevate rispetto ai lavoratori con livello di istruzione inferiore. Usando dati provenienti dal Current Population Survey del marzo 2013, si stima la relazione tra la retribuzione oraria e il numero di anni di istruzione per un campione di 2829 lavoratori a tempo pieno negli Stati Uniti di età compresa tra 29 e 30 anni, con anni di istruzione compresi tra 6 e 18:

$$\widehat{Earnings} = -7,29 + 1,93 * YearsEducation, \quad R^2 = 0,162 \quad SER = 10,29$$

(1,10) (0,08)

- Un lavoratore di 30 anni selezionato casualmente ha un livello di istruzione pari a 16 anni (ovvero è un laureato). Qual è la retribuzione oraria attesa del lavoratore?

- b) Un diplomato (con 12 anni di istruzione) considera la possibilità di iscriversi all'università per ottenere una laurea biennale. Di quanto ci si attende che cresca la sua retribuzione oraria media?
- c) Un consulente scolastico di una scuola superiore dice a uno studente che, in media, i laureati guadagnano 10\$ all'ora in più rispetto ai diplomati (dunque, il Delta in anni di Education è di 4 e $\beta_1 = 2,50$). Questa affermazione è coerente con l'evidenza della regressione? Quale intervallo di valori è coerente con l'evidenza della regressione (Suggerimento: l'intervallo di confidenza dev'essere calcolato per $\Delta * \beta_1$)?

CAPITOLI 6-7 “Regressione Lineare con Regressori Multipli – Verifica di Ipotesi e Intervalli di Confidenza”

DOMANDA 1.

Cosa si intende per distorsione da variabile omessa? Si fornisca una definizione e si definiscano le condizioni che devono verificarsi.

Quale assunzione dei minimi quadrati viene meno se vi è distorsione da variabili omesse?

DOMANDA 2.

Che cos'è la trappola della variabile dummy e che relazione ha con la multicollinearità dei regressori?

Qual è la soluzione per questa forma di multicollinearità?

DOMANDA 3.

Che differenza c'è tra la collinearità imperfetta dei regressori e la collinearità perfetta? Si confrontino le soluzioni per questi due problemi legati alla regressione multipla.

DOMANDA 4.

Elencare le assunzioni dei minimi quadrati nel caso di regressione con un singolo regressore.

Quali sono le assunzioni dei minimi quadrati nel caso di regressione multipla?

I seguenti esercizi fanno riferimento alla tabella delle regressioni stimate (*Tabella 6.2*) usando dati per il 2012 tratti dal Current Population Survey. I dati consistono di informazioni su 7440 lavoratori a tempo pieno occupati per tutto l'anno. Il titolo di studio più elevato di un lavoratore è il diploma o la laurea. L'età del lavoratore è compresa tra 25 e 34 anni. I dati contengono anche informazioni sulla regione dove la persona vive, lo stato civile e il numero di figli. Per gli scopi di questi esercizi siano:

- *AHE*=retribuzione oraria media (in dollari 2012)
- *College*=variabile binaria (1 se università, 0 se scuola superiore)
- *Female*=variabile binaria (1 se femmina, 0 se maschio)
- *Age*=età (in anni)
- *Ntheast*=variabile binaria (1 se regione = Nord-Est, 0 altrimenti)
- *Midwest*=variabile binaria (1 se regione = Centro-Ovest, 0 altrimenti)
- *South*=variabile binaria (1 se regione = Sud, 0 altrimenti)
- *West*=variabile binaria (1 se regione = Ovest, 0 altrimenti).

Tabella 6.2 Risultati di regressioni della retribuzione oraria media sulle variabili binarie genere e istruzione e su altre caratteristiche usando i dati del 2012 tratti dal *Current Population Survey*.

Variabile dipendente: retribuzione oraria media (AHE, Average Hourly Earnings)			
Regressore	(1)	(2)	(3)
College (X_1)	8,31	8,32	8,34
Female (X_2)	-3,85	-3,81	-3,80
Age (X_3)		0,51	0,52
Northeast (X_4)			0,18
Midwest (X_5)			-1,23
South (X_6)			-0,43
Intercetta	17,02	1,87	2,05
Statistiche sintetiche			
SER	9,79	9,68	9,67
R^2	0,162	0,180	0,182
n	7440	7440	7440

DOMANDA 5.

Usando i risultati nella colonna (1):

- I lavoratori laureati, in media, guadagnano più dei lavoratori con solo il diploma? Quanto di più?
- In media, gli uomini guadagnano più delle donne? Quanto di più?

DOMANDA 6.

Usando i risultati della regressione nella colonna (2):

- L'età è una determinante importante delle retribuzioni? Si argomenti la risposta.
- Sally è una donna laureata di 29 anni. Betsy è una donna laureata di 34 anni. Si predicano le retribuzioni di Sally e Betty.

DOMANDA 7.

Usando i risultati della regressione (3):

- Ci sono importanti differenze regionali?
- Perché il regressore *West* è omissso dalla regressione? Che cosa accadrebbe se venisse incluso?
- Quale statistica utile per effettuare un test di ipotesi congiunte sui coefficienti delle regressioni manca? Quale tipo di informazione potrebbe fornirci? Si argomenti la risposta.