BANCA D'ITALIA
EUROSISTEMA

CONCORSO PER L'ASSUNZIONE DI 5 ESPERTI LAUREATI CON ORIENTAMENTO NELLE DISCIPLINE STATISTICHE
(Bando 15/09/2020 – Lett. B)

Testo n. 3

STATISTICA E PROBABILITÀ

Due quesiti a scelta tra tre proposti dalla Commissione

QUESITO N. 1

1. Si consideri un campione casuale $(X_1, \ldots\ldots, X_n)$ e uno stimatore $T(X) = T(X_1, \ldots\ldots, X_n)$ di un parametro θ. La candidata/il candidato esponga brevemente le proprietà di correttezza, consistenza ed efficienza di uno stimatore e le eventuali relazioni tra tali proprietà. Si esponga inoltre il risultato del teorema di Cramér – Rao evidenziando le ipotesi alla base del teorema.

2. Si consideri un campione casuale estratto in modo indipendente da una popolazione con funzione di densità:

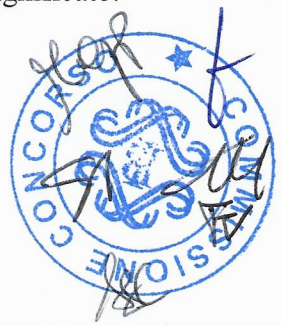
$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x \leq \theta, 0 < \theta < \infty \\ 0, & altrove \end{cases}$$

Si mostri che lo stimatore $X_{(n)} = massimo\ elemento\ campionario$ è uno stimatore distorto ma consistente del parametro θ.

3. Si individui l'estremo inferiore della disuguaglianza di Cramér – Rao e si mostri che lo stimatore corretto $Z(X) = \frac{(n+1)}{n} X_{(n)}$ ha varianza inferiore a tale limite, spiegandone il motivo.

QUESITO N. 2

1. Sia $A_1, \ldots, A_j, \ldots, A_r$ un insieme finito di eventi incompatibili e necessari, la cui unione $\bigcup_{j=1}^{r} A_j = \Omega$ restituisce l'evento certo. Sia E un evento tale che $E \subset \bigcup_{j=1}^{r} A_j$ e $P(E) \neq 0$. Si illustri il teorema di Bayes e si dia una spiegazione, fornendo un esempio, del suo significato.



2. In un'urna ci sono y palline, di cui x bianche e y-x nere. In una seconda urna ci sono ancora y palline, ma quelle bianche sono y-x. Si prende a caso una pallina dalla prima urna e la si mette nella seconda. Poi, dalla seconda urna si prende a caso una pallina e la si mette nella prima. Si calcoli la distribuzione di probabilità dell'evento A_r = *"nella prima urna ci sono alla fine r palline bianche"*.

3. Si abbiano ora tre urne, ciascuna contenenti due palline bianche e due nere. Si estrae una pallina dalla prima urna e la si mette nella seconda. Da questa si estrae una pallina e la si mette nella terza e infine si estrae una pallina dalla terza mettendola nella prima. Trovare la distribuzione della variabile aleatoria doppia *(X, Y)="numero di palline bianche finali nella prima e nella seconda urna"*.

QUESITO N. 3

Si supponga di voler verificare l'efficacia di un determinato sonnifero e che a questo scopo si sia deciso di ricorrere a un esperimento coinvolgendo 10 individui, raggruppati casualmente in due gruppi di numerosità pari a 5. Al primo gruppo viene somministrato il sonnifero per un periodo di una settimana (gruppo A), mentre al secondo viene somministrato un placebo, sempre per un periodo di una settimana (gruppo B). Prima dell'avvio dell'esperimento, il livello d'insonnia nei due gruppi era lo stesso. Trascorso il periodo di assunzione del sonnifero e del placebo, ai 10 individui è stato chiesto di esprimere un voto circa la qualità del sonno (un voto pari a zero implica assenza di miglioramento). I voti sono i seguenti:

Gruppo A: 10, 1, 2, 0, 2

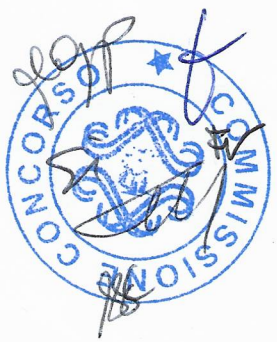
Gruppo B: 11, 10, 5, 7, 3

La candidata/il candidato:

1. sottoponga a verifica con un livello di significatività del 5% l'ipotesi che le mediane siano uguali contro l'ipotesi alternativa bidirezionale di mediane diverse applicando il test Mann–Whitney U;

2. descriva le condizioni che devono essere rispettate per applicare il test e illustri come si può modificare la valutazione dei risultati nel caso in cui la numerosità dei campioni sia più elevata (n>30);

3. illustri brevemente i principali test non parametrici utilizzabili nel caso di variabili almeno ordinabili.



ECONOMETRIA E *STATISTICAL LEARNING*

Un quesito a scelta tra due proposti dalla Commissione

QUESITO N. 4

1. Si consideri una variabile risposta y i cui valori $y_1, \ldots, y_n$ rappresentano il risultato di conteggi e la cui distribuzione è caratterizzata da una certa media e da una certa varianza. L'obiettivo dell'analisi è prevedere (spiegare) il numero medio di conteggi sulla base di un predittore lineare:

$$X_i'\beta = \sum_{j=1}^{p} x_{ij}\beta_j\,; \quad i = 1, \ldots, n$$

A tal fine si stima il seguente modello di regressione:

$$E(y_i|X_i'\beta) = \mu_i = exp(X_i'\beta)$$

dove y è il vettore della variabile risposta osservata su un campione di n individui, X_i un vettore di p predittori di qualsiasi natura relativo all'i-esimo individuo.

La candidata/il candidato:

1.a espliciti le principali caratteristiche del modello di cui sopra e illustri eventuali differenze con il modello di regressione lineare;

1.b inquadri il modello stimato all'interno dei modelli lineari generalizzati e ne espliciti la funzione *link*.

2. Su un campione di 915 studenti di dottorato è stato stimato un modello log-lineare per prevedere il numero di pubblicazioni scientifiche pubblicate dagli studenti durante i tre anni di corso di dottorato.

```
. glm art fem mar kid5 phd ment, family(poisson) nolog

Generalized linear models                         No. of obs      =        915
Optimization     : ML                             Residual df     =        909
                                                  Scale parameter =          1
Deviance         =  1634.370984                   (1/df) Deviance =   1.797988
Pearson          =   1662.54655                   (1/df) Pearson  =   1.828984

------------------------------------------------------------------------------
             |                 OIM
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       donna |  -.2245942   .0546138    -4.11   0.000    -.3316352   -.1175532
     sposato |   .1552434   .0613747     2.53   0.011     .0349512    .2755356
      figli5 |  -.1848827   .0401272    -4.61   0.000    -.2635305   -.1062349
         phd |   .0128226   .0263972     0.49   0.627     -.038915    .0645601
        ment |   .0255427   .0020061    12.73   0.000     .0216109    .0294746
       _cons |   .3046168   .1029822     2.96   0.003     .1027755    .5064581
------------------------------------------------------------------------------
```

dove:

- la variabile risposta "art" rappresenta il numero di articoli pubblicati dallo studente;
- "donna" è una variabile binaria che assume valore 1 se l'individuo intervistato è donna e 0 se è uomo;
- "sposato" è una variabile binaria che assume valore 1 se l'individuo intervistato è sposato/convivente e 0 altrimenti;
- "figli5" è una variabile numerica che indica il numero di figli di età compresa tra 0 e 5 anni;
- "phd" è una variabile numerica che rappresenta un indice di prestigio del dipartimento di afferenza, definito su una scala da 0 a 5;
- "ment" è una variabile numerica che rappresenta il numero di articoli pubblicati dal mentore dello studente durante i 3 anni del corso di dottorato.

La candidata/il candidato:

2.a commenti l'output del modello stimato, con particolare attenzione all'interpretazione dei coefficienti nello spiegare l'andamento della variabile "numero degli articoli pubblicati dallo studente";

2.b sapendo che esiste forte sovradispersione e che la stima del fattore di dispersione è pari a circa 1.35, discuta come si modificherebbero i risultati se tenessimo conto della sovradispersione.

3. La candidata/il candidato discuta l'eventualità di includere nel modello di regressione indicato al punto 1 una variabile di esposizione al rischio *exposure*.

QUESITO N. 5

1. Si consideri il seguente modello per dati *longitudinali*:

$$y_{i,t} = X'_{i,t}\beta + \mu_i + \epsilon_{i,t} \qquad t = 1, \dots, T \quad i = 1, \dots, N$$

dove $y_{i,t}$ rappresenta la variabile di risposta per l'unità *i*-esima osservata al tempo *t*, $X_{i,t}$ è un vettore di K predittori per l'unità *i*-esima osservata al tempo *t*, μ_i l'effetto individuale e $\epsilon_{i,t}$ il termine di errore idiosincratico.

1.a La candidata/il candidato dimostri come si ottiene lo stimatore *fixed effects* e discuta l'assunzione di esogeneità stretta;

1.b Quale assunzione aggiuntiva sull'effetto individuale caratterizza lo stimatore *random effects*? Quali vantaggi offre lo stimatore *random effects* rispetto a quello *fixed effects*?

2. Un ricercatore è interessato a studiare i comportamenti di risparmio delle famiglie italiane. Utilizzando un panel bilanciato di 2139 famiglie intervistate per 4 anni consecutivi, ha stimato il seguente modello, sia con lo stimatore *fixed effects* (FE) sia con quello *random effects* (RE):

	FE risparmi (log)	RE risparmi (log)
redditi (log)	0.8273***	1.0750***
	(0.0396)	(0.0308)
attivita' finanziarie (log)	0.0386***	0.1079***
	(0.0146)	(0.0106)
immobili (log)	0.0495***	0.1290***
	(0.0150)	(0.0078)
mutui (log)	-0.0482***	-0.0437***
	(0.0133)	(0.0095)
costante	-1.6655***	-5.4336***
	(0.4252)	(0.2956)
R-quadro: within	0.082	0.078
R-quadro: between	0.487	0.498
R-quadro: overall	0.259	0.272

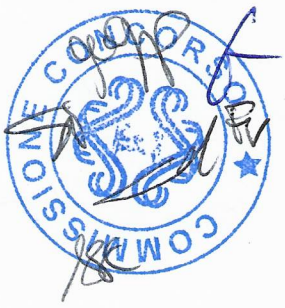
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

La variabile dipendente è data dai risparmi annuali della famiglia in euro (trasformati in logaritmo). I predittori, espressi in euro e trasformati in logaritmo, sono i seguenti: redditi (annuali), attività finanziarie al valore di mercato a fine anno, valore degli immobili a fine anno, importo dei mutui a fine anno.

2.a La candidata/il candidato commenti i risultati della prima colonna (FE) e in particolare il coefficiente sulla variabile *redditi (log)*. Descriva poi l'R-quadro *between* e quello *within*. Se si volesse testare l'importanza di includere gli effetti fissi nel modello come si dovrebbe procedere?

2.b La candidata/il candidato discuta brevemente le differenze tra i coefficienti delle due colonne. Se si effettuasse un test di Hausman ottenendo un p-value pari a 0, cosa se ne evincerebbe? Il candidato illustri sinteticamente la logica del test.

3. La candidata/il candidato spieghi in quali circostanze lo stimatore *first difference* sia preferibile a quello *fixed effects*.



METODI DI CAMPIONAMENTO

Un quesito a scelta tra due proposti dalla Commissione

QUESITO N. 6

Si intende implementare un disegno di campionamento per stimare il reddito medio delle famiglie che vivono in un quartiere suddiviso in 10 isolati. La popolazione del quartiere è composta da 1.000 famiglie. Si decide di selezionare tre isolati attraverso un campionamento casuale semplice senza reintroduzione e di intervistare tutte le famiglie che vivono nei tre isolati. I risultati dell'indagine sono riassunti nella tavola che segue.

Isolato	Numero di famiglie	Reddito medio campionario
1	120	18
2	100	20
3	50	22

1. Si descriva brevemente il disegno campionario adottato discutendone vantaggi e svantaggi. Si indichino, inoltre, le probabilità di inclusione del primo e del secondo ordine, e si stimi il reddito medio famigliare utilizzando lo stimatore di Horvitz-Thompson.

2. Si spieghi quali limiti può avere la strategia campionaria adottata in relazione al fatto che gli isolati hanno numerosità diversa e si indichino possibili strategie di campionamento alternative al fine di ottenere una stima più accurata del reddito medio.

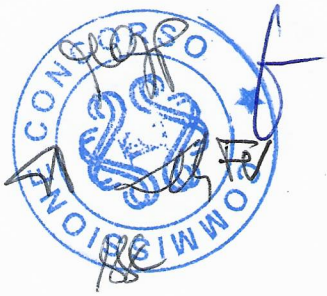
3. Si supponga, infine, che non tutte le famiglie intervistate abbiano fornito il valore del proprio reddito famigliare ma che tutte abbiano invece riportato il titolo di studio del capofamiglia. In che modo è possibile impiegare questa informazione per migliorare la qualità della stima campionaria del reddito famigliare?

QUESITO N. 7

Si supponga di voler realizzare un'indagine campionaria, con campionamento casuale semplice senza reintroduzione, per stimare la spesa media mensile in generi alimentari consumati fuori casa (ad es. al ristorante) di una popolazione di 1.000 famiglie. Da una precedente indagine è emerso che la media e la varianza della spesa in generi alimentari consumati fuori casa sono pari a 70 e 140, rispettivamente.

1. Quale numerosità campionaria è necessaria per garantire un coefficiente di variazione non superiore a 0,03 per la stima della spesa media di generi alimentari consumati fuori casa?

2. Si supponga di conoscere il numero medio di componenti per famiglia per l'intera popolazione delle famiglie. Ipotizzando un rapporto di proporzionalità approssimata tra la spesa famigliare



per generi alimentari consumati fuori casa e il numero di componenti, si illustri quale stimatore si può adottare per migliorare la stima e se ne discutano le proprietà.

3. Si ipotizzi che l'indagine sia condotta tramite *web* e che il campione di famiglie sia estratto da un elenco di indirizzi email che non contiene tutta la popolazione di riferimento. Quali potrebbero essere le conseguenze derivanti dall'uso di tale elenco sulla stima della spesa media in generi alimentari consumati fuori casa?

PROVA IN LINGUA INGLESE

Do you think the economy will revive immediately and people will go back to their pre-pandemic spending habits when the restrictions imposed by the pandemic are lifted? Why or why not?

Table A11 Table of critical Values for Mann–Whitney *U* Statistic (continued)

(Two–Tailed .01 Values)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																			0	0
3									0	0	0	1	1	1	2	2	2	2	3	3
4						0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5					0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6				0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7				0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8				1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9			0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10			0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11			0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12			1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13			1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14			1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15			2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16			2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17			2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18			2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19		0	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20		0	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

(One–Tailed .01 Values)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2													0	0	0	0	0	0	1	1
3							0	0	1	1	1	2	2	2	3	3	4	4	4	5
4					0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6				1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7			0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8			0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9			1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
10			1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
11			1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
12			2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
13		0	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
14		0	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
15		0	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
16		0	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
17		0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
18		0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
19		1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
20		1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114

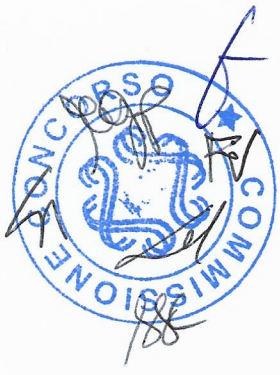


Table A11 Table of Critical Values for Mann–Whitney *U* Statistic

(Two-Tailed .05 Values)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

(One-Tailed .05 Values)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																			0	0
2					0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3			0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4			0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5		0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6		0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7		0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
8		1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
9		1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
10		1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
11		1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
12		2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
13		2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
14		2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
15		3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
16		3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
17		3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
18		4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138