Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in

Statistica e Tecnologie Informatiche

RELAZIONE FINALE

SULLA STIMA JACKKNIFE DELLA VARIANZA DI UNO STIMATORE PER L'AREA SOTTO LA CURVA ROC

Relatore Prof. Gianfranco Adimari
Dipartimento di Scienze Statistiche

Laureando: Stefano Mussi
Matricola N. 1010128

Anno Accademico 2012/2013

Riassunto

La curva ROC è di solito utilizzata per valutare la capacità discriminatoria di un test diagnostico, cioè di un test che dovrebbe dividere la popolazione in positivi e negativi, in particolare in campo medico i pazienti malati da quelli sani. Adimari e Chiogna (2012) hanno proposto uno stimatore non-parametrico per trattare la distorsione causata dai pazienti in cui non si è verificata l'effettiva presenza o meno della malattia, assegnando loro, come stima della probabilità di essere malati, la media dei valori relativi allo stato (malato/sano) dei k vicini-più-vicini verificati (stimatore-KNN). Inoltre gli stessi autori hanno dimostrato come il relativo stimatore dell'area sotto la curva ROC (AUC), sia consistente e normalmente distribuito. In questa tesi si studierà la stima Jackknife della varianza di tale stimatore confrontandola col valore calcolato via Monte Carlo, per poi applicare il risultato a dei dati reali con lo scopo di calcolare degli intervalli di confidenza per l'AUC.

Indice

Capitolo 1

La curva ROC

1.1 I test diagnostici

Spesso si ha la necessità di dividere un gruppo di unità in due sottogruppi distinti in base alla presenza o meno di una determinata caratteristica. I test che si occupano di svolgere tale funzione vengono denominati diagnostici, in particolare se i risultati del test utilizzato sono continui si parla di test diagnostici su scala continua. In campo medico si è interessati a determinare la presenza o meno di una patologia su dei pazienti.

Spesso se il risultato di un test diagnostico è maggiore di un valore fissato a priori, denominato valore di cut-off (c), allora è considerato positivo, in campo medico malato, mentre se è minore allora è considerato negativo e il paziente quindi sano. Un test che separa i pazienti nei due gruppi senza compiere errori di classificazione è detto test diagnostico gold standard (GS) e viene utilizzato come riferimento per verificare l'efficienza dei test che si prefiggono di effettuare la stessa analisi. Vengono proposti test con la stessa funzione perchè i GS sono spesso o dispendiosi o invasivi nei confronti dei pazienti e per questo si preferisce sottoporli solo a una parte del campione. I soggetti a cui vengono effettuati i GS vengono scelti in base ai singoli risultati del test alternativi o in base a loro particolari caratteristiche personali. Se la scelta di questo sotto-

gruppo è indipendente dal loro stato effettivo di salute allora vengono soddisfatte le assunzioni MAR (missing at random).

Per ogni paziente si hanno quindi le seguenti informazioni:

- X, il vettore delle covariate osservate sul soggetto;
- T, il valore del test;
- V, vale 1 se lo stato di salute è stato verificato tramite GS, 0 altrimenti;
- D, eventuale indicatore dello stato di salute e vale 0 se il soggetto è sano, 1 se malato.

Se ci si attiene alle assunzioni MAR si introducono due nuove probabilità, $\pi = Pr(V = 1|T, X)$, cioè la probabilità che il soggetto sia verificato condizionatamente al risultato del test e alle covariate, e $\rho = Pr(D = 1|T, X)$, cioè la probabilità di risultare positivo ad un eventuale GS.

1.2 Sensitività, specificità e AUC

Dal confronto del GS con un altro test che ha lo stesso scopo esplorativo si può dividere la popolazione in quattro gruppi, i TP, che vengono classificati positivi da entrambi i test, i TN, classificati negativi da entrambi i test, i FP, negativi per il GS e positivi per l'altro test e infine i FN che sono quelli considerati positivi dal GS ma non dall'altro test. Da queste quattro quantità vengono delineati due valori che possono essere utilizzati per determinare la capacità del test. Essi sono la sensitività, che è la probabilità che il test classifichi positivo un soggetto realmente malato, e la specificità, che invece è la probabilità che lo stesso test classifichi negativo un soggetto sano. Sono quindi espresse come:

$$SE(c) = \frac{TP}{TP + FP} \quad SP(c) = \frac{TN}{TN + FN}$$

Sulla base di un campione di n unità statistiche, se tutte le unità sono verificate ($V_i = 1 \quad \forall i = 1, 2, ;n$), allora la sensitività e la specificità

possono essere stimate come:

$$\widehat{SE}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c) D_i}{\sum_{i=1}^{n} D_i} \tag{1.1}$$

$$\widehat{SP}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)(1 - D_i)}{\sum_{i=1}^{n} (1 - D_i)} \tag{1.2}$$

La specificita e la sensitività dipendono dal punto di cut-off e sono tra loro inversamente proporzionali al variare di questo valore. Questo permette di poter fissare c in base ai livelli di specificità e di sensitività desiderati. Se per esempio si analizza una patologia ad alto rischio si preferisce avere un livello di sensitività molto alto per non rischiare di trascurare un soggetto a rischio, mentre se la cura è costosa e la malattia non molto grave ci si focalizza sulla specificità.

Per avere un giudizio complessivo sul test è stata introdotta la curva ROC (Receiver Operating Characteristic). Essa è stata utilizzata per la prima volta durante la II guerra mondiale dagli ingegnieri elettronici americani per migliorare la capacità di previsione dei radar, oggi invece è molto popolare in campo medico. La curva poggia su un piano cartesiano, ha sull'asse delle ascisse *1- la specificità* mentre sulle ordinate *la sensitività*. Per disegnarla si fa variare c e in corrispondenza di ogni suo valore considerato si calcolano la sensitività e la specificità. L'insieme dei punti calcolati se riportati graficamente formano la curva ROC.

L'area sotto questa curva (AUC) varia tra 0.5 e 1 (in realtà se il test assegna ai negativi valori più alti rispetto ai positivi, essa varia fra 0 e 0.5, problema facilmente risolvibile cambiando di segno i valori del test) e più si avvicina all'unità, maggiore è la capacità del test di discriminare le due popolazioni. L'AUC è in altri termini la probabilità che il test assegni ad un soggetto sano un valore minore di un qualsiasi malato, quindi $Pr(T_i < T_j \,|\, D_i = 0 \vee D_j = 1)$. Questa probabilità può essere stimata (in maniera non parametrica) contando quante volte il valore del test per i sani nel campione precede quello dei malati e dividendo il tutto per il prodotto delle numerosità dei due gruppi (sani e malati nel

campione). Se tutte le unità sono verificate ($V_i = 1 \quad \forall i = 1, 2, \ldots, n$):

$$\widehat{AUC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I(T_i > T_j) D_i (1 - D_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} D_i (1 - D_j)} \tag{1.3}$$

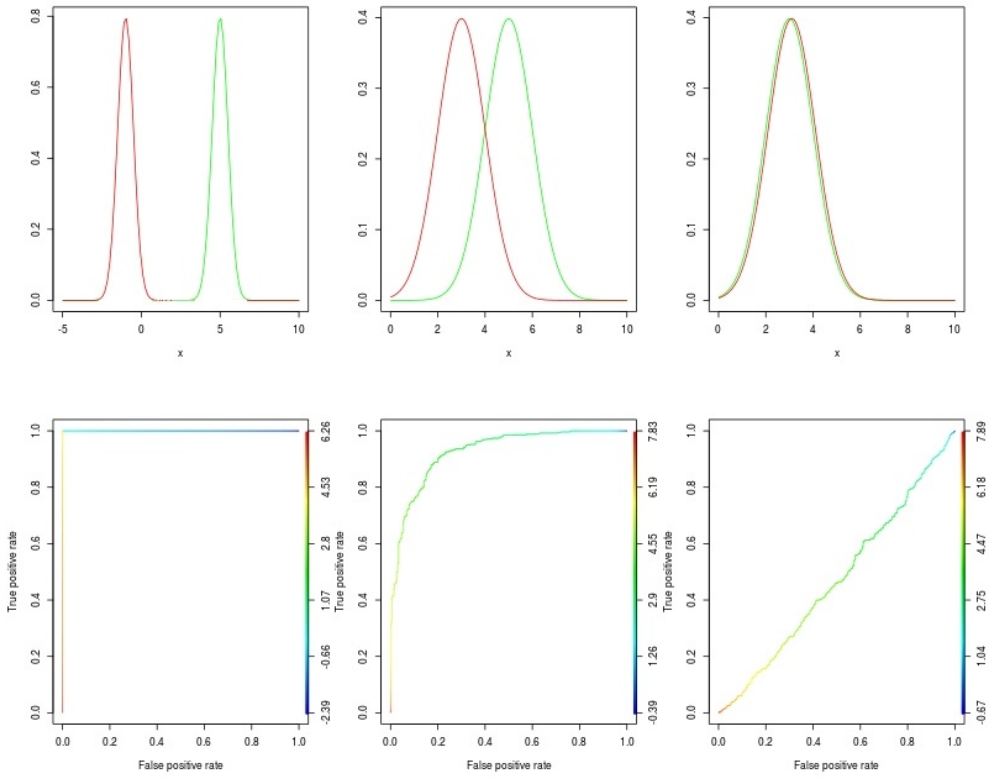


Fig. 1.1: Nella prima riga sono riportate i risultati di tre test diversi divisi per sani, in verde, e malati, in rosso. Il test più a sinistra è un GS. Nella seconda riga sono riportate le rispettive curve ROC

1.3 Verifica parziale

Se solo a una parte dei soggetti sotto studio è stato vericato l'effettivo stato di salute per i motivi sopraccitati, sono stati proposti in letteratura alcuni stimatori che si basano sulle assunzioni MAR. I principali per la specificità, sensitività e AUC sono:

- Il Full Imputation (FI) che prevede la specificazione di un modello di regressione per ρ, da cui ottenere stime $\hat{\rho}_i$ usando solo i dati dei

soggetti verificati:

$$\widehat{SE}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\hat{\rho}_i}{\sum_{i=1}^{n} \hat{\rho}_i}$$

$$\widehat{SP}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)(1-\hat{\rho}_i)}{\sum_{i=1}^{n} (1-\hat{\rho}_i)}$$

$$\widehat{AUC} = \frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} I(T_i > T_j)\hat{\rho}_i(1-\hat{\rho}_j)}{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} \hat{\rho}_i(1-\hat{\rho}_j)}$$

- Il Mean Score Imputation è un'altro possibile approccio che, a differenza del FI, prevede di inputare solo i dati effettivamente mancanti:

$$\widehat{SE}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (1-V_i)\hat{\rho}_i\}}{\sum_{i=1}^{n} \{V_i D_i + (1-V_i)\hat{\rho}_i\}}$$

$$\widehat{SP}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_i)\}}{\sum_{i=1}^{n} \{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_i)\}}$$

$$\widehat{AUC} = \frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} I(T_i > T_j)\{V_i D_i + (1-V_i)\hat{\rho}_i\}(1-\{V_j D_j + (1-V_j)\hat{\rho}_j\})}{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} \{V_i D_i + (1-V_i)\hat{\rho}_i\}(1-\{V_j D_j + (1-V_j)\hat{\rho}_j\})}$$

- L'Inverse Probability Weighting (IPW) prevede la specificazione di un modello di regressione per π:

$$\widehat{SE}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c) V_i D_i \hat{\pi}_i^{-1}}{\sum_{i=1}^{n} V_i D_i \hat{\pi}_i^{-1}}$$

$$\widehat{SP}(c) = \frac{\sum_{i=1}^{n} I(T_i < c) V_i (1-D_i) \hat{\pi}_i^{-1}}{\sum_{i=1}^{n} V_i (1-D_i) \hat{\pi}_i^{-1}}$$

$$\widehat{AUC} = \frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} I(T_i > T_j) V_i D_i \hat{\pi}_i^{-1} V_j (1-D_j) \hat{\pi}_j^{-1}}{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} V_i D_i \hat{\pi}_i^{-1} V_j (1-D_j) \hat{\pi}_j^{-1}}$$

- Il Semi-Parametric Efficient (SPE) che è consistente se $\hat{\pi}_i$ o $\hat{\rho}_i$ sono stime consistenti:

$$\widehat{SE}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (\hat{\pi}_i - V_i)\hat{\rho}_i\}\hat{\pi}_i^{-1}}{\sum_{i=1}^{n} \{V_i D_i + (\hat{\pi}_i - V_i)\hat{\rho}_i\}\hat{\pi}_i}$$

$$\widehat{SP}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)\{V_i(1-D_i) + (\hat{\pi}_i - V_i)(1-\hat{\rho}_i)\}\hat{\pi}_i^{-1}}{\sum_{i=1}^{n} \{V_i(1-D_i) + (\hat{\pi}_i - V_i)(1-\hat{\rho}_i)\}\hat{\pi}_i^{-1}}$$

$$\widehat{AUC} = \frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} \big\{ I(T_i > T_j)\{V_i D_i + (\hat{\pi}_i - V_i)\hat{\rho}_i\}\hat{\pi}_i^{-1}}{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} \big\{ \{V_i D_i + (\hat{\pi}_i - V_i)\hat{\rho}_i\}\hat{\pi}_i^{-1}} \cdot$$

$$\cdot \frac{\{V_j(1-D_j) + (\hat{\pi}_j - V_j)(1-\hat{\rho}_j)\}\hat{\pi}_j^{-1}\big\}}{\{V_j(1-D_j) + (\hat{\pi}_j - V_j)(1-\hat{\rho}_j)\}\hat{\pi}_j^{-1}\big\}}$$

Le probalilità $\hat{\pi}_i$ e $\hat{\rho}_i$ vengono stimate attraverso un modello di regressione. Se questo modello non è correttamente specificato la bontà di previsione degli stimatori è fortemente compromessa. Per evitare questo problema in Adimari e Chiogna (2012) viene proposto, sotto le assunzioni MAR, uno stimatore completamente non parametrico per i $\hat{\rho}_i$. Adimari e Chiogna propongono di stimare la probabilità che l'i-esimo soggetto sia malato utilizzando la media dei K valori di D dei soggetti verificati "più vicini":

$$\hat{\rho}_i = \frac{1}{K}\sum_{j=1}^{n} D_{i(j)} \tag{1.4}$$

dove $\{(Y_{i(j)}, D_{i(j)} : V_{i(j)} = 1, j = 1, \ldots, K\}$ è un insieme di K dati e $Y_{i(j)}$ indica la j-esima osservazione verificata più vicina a $Y_i = (T_i, X_i)^T$. Per il calcolo dei vicini più vicini viene qua scelta la distanza euclidea. Da questa quantità vengono delineati i seguenti stimatori consistenti e asintoticamente normali sotto le assunzioni MAR:

$$\hat{\theta}_1 \;=\; Pr(D=1) \;=\; \frac{1}{n}\sum_{i=1}^{n}\{V_i D_i + (1-V_i)\hat{\rho}_{Ki}\}$$

$$\hat{\theta}_2 \;=\; Pr(T_i \geq c, D=1) \;=\; \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (1-V_i)\hat{\rho}_{Ki}\}$$

$$\hat{\theta}_3 \;=\; Pr(T_i \geq c, D=0) \;=\; \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)\{V_i(1-D_i)+(1-V_i)(1-\hat{\rho}_{Ki})\}$$

Di conseguenza, gli stimatori della sensitività $\widehat{SE}(c) = \frac{\hat{\theta}_2}{\hat{\theta}_1}$ e la specificità $\widehat{SP}(c) = 1 - \frac{\hat{\theta}_3}{1-\hat{\theta}_1}$ sono anch'essi consistenti e asintoticamente normali. Gli stessi autori hanno anche derivato uno stimatore per l'AUC che gode delle stesse proprietà. Viene dunque presentata una nuova forma non parametrica degli stimatori MSI.

$$\widehat{SE}_{KNN}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (1-V_i)\hat{\rho}_{Ki}\}}{\sum_{i=1}^{n}\{V_i D_i + (1-V_i)\hat{\rho}_{Ki}\}} \tag{1.5}$$

$$\widehat{SP}_{KNN}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_{Ki})\}}{\sum_{i=1}^{n}\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_{Ki})\}} \tag{1.6}$$

$$\widehat{AUC}_{KNN} = \frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} I(T_i > T_j)\tilde{D}_i(1-\tilde{D}_j)}{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} \tilde{D}_i(1-\tilde{D}_j)} \quad \tilde{D}_i = V_i D_i + (1-V_i)\hat{\rho}_{Ki} \tag{1.7}$$

Il comportamento di quest'ultimo stimatore è già stato studiato in una precedente tesi, Zamengo B. (2013). In particolare Zamengo analizza la proprietà dello stimatore Bootstrap per la stima della varianza dello stimatore. In questa tesi si fa lo stesso studio via Jackknife. Questo permette di ridurre i tempi e quindi dà la possibilità di aumentare la numerosità campionaria e di estendere lo studio a un secondo scenario non trattato nella sopraccitata tesi.

Capitolo 2

Il metodo Jackknife

2.1 Il ricampionamento

I metodi di ricampionamento si sono sviluppati recentemente grazie all'aumento della potenza computazionale e vengono usati principalmente in ambito non parametrico. In campo parametrico si ha che $F(x)$, la legge di probabiltà della v.c X da cui deriva il campione, è nota a meno del valore di θ. Inoltre $\hat{\theta}$, uno stimatore per θ ignoto, ha anch'esso una legge di probabilità nota (a meno di θ) poichè dipende dalla $F(x)$.

Si è invece in campo non parametrico se l'insufficienza di informazioni su X non permettono di ipotizzare una distribuzione nota per $F(x)$, o se la complessità dello stimatore $\hat{\theta}$ non permette il riferimento a una legge di probabilità definita. Per affrontare questi problemi si può fare riferimento ai risultati asintotici, cioè quei risultati che valgono per n infinito, ovvero nella pratica per n sufficientemente grande.

Un'altra strada è quella di utilizzare i metodi di ricampionamento, cioè quei metodi che, a causa della carenza di informazioni su X, creano una serie di nuovi campioni sulla base del campione originale. Da questo vengono infatti estratte le osservazioni che andranno a formare i nuovi campioni. I metodi di ricampionamento quindi vengono usati per valutare l'accuratezza di uno stimatore in ambito non parametrico. Infatti una volta ottenuto $\hat{\theta}$ si è interessati a valutare la sua distorsione e la disper-

sione dello stimatore intorno al parametro che nel caso di uno stimatore non distorto è la varianza. Lo studio della varianza di uno stimatore è l'obiettivo di questa tesi. I più noti metodi di ricampionamento sono il Bootstrap, qua non trattato, e il Jackknife.

2.2 Il Jackknife

Il metodo Jackknife è stato proposto nel 1949 da M. H. Quenouille che, a causa della scarsa potenza computazionale dell'epoca, creò un algoritmo che richiede un numero fissato di conti. In inglese il termine jackknife indica il coltellino serramanico, infatti l'idea principale di questo metodo è di "tagliare" ogni volta un'osservazione diversa dal campione originale e ogni volta ristimare il parametro d'interesse. La stima verrà confrontata con la stessa calcolata sul campione originale. Nel dettaglio l'algoritmo si basa su questi passi:

- Si creano n campioni Jackknife togliendo ogni volta l'osservazione i-esima, quindi l'i-esimo campione è formato da $[x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$
- Per ogni campione si stima il parametro θ. Si hanno dunque n stime di θ: $[\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_n]$
- La media di queste stime $\bar{\hat{\theta}}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i$ è uno stimatore di θ
- Si calcola $\hat{\theta}$ sul campione originario
- I Pseudo-valori vengono definiti come $\theta_i^* = n\hat{\theta} - (n-1)\hat{\theta}_i$
- La media di questi valori $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \theta_i^*$ è detta stima Jackknife corretta.
- Infine la stima Jackknife della varianza è definita come $V_{JK} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\theta_i^* - \tilde{\theta})$

Capitolo 3

Simulazione

Per studiare la capacità del metodo Jackknife per la stima della varianza dello stimatore $\widehat{AUC}_{KNN}$ viene ripreso lo schema di simulazione presente in Adimari e Chiogna (2012). Vengono presi in considerazione tutti e due gli scenari fissati in quel lavoro. In entrambi vengono generate 5000 replicazioni Monte Carlo. Per ognuna di esse viene utilizzato il metodo Jackknife per la stima della varianza e della media di $\widehat{AUC}_{KNN}$, viene inoltre calcolato $\widehat{AUC}_{KNN}$ su tutte le osservazioni con la (1.7) e lo stimatore full (1.3) ottenuto utilizzando tutte le informazioni sulle unità. Finite le 5000 replicazioni è fatta la media di tutti i valori e inoltre viene calcolata la varianza di $\widehat{AUC}_{KNN}$ (varianza Monte Carlo calcolata su tutte le 5000 replicazioni). Questa verrà poi confrontata con la media delle stime Jackknife della varianza di $\widehat{AUC}_{KNN}$. Per ogni replicazione vengono generate quattro normali con la seguente distribuzione:

$Z_1 \sim Z_2 \sim N(0, 0.5)$

$e_1 \sim e_2 \sim N(0, 0.25)$

Da queste vengono generati il valore del test T, le covariate X e lo stato di salute:

$$T = h(Z_1; Z_2) + e_1;$$
$$X = f(Z_1; Z_2) + e_2;$$
$$D = I(g(Z_1; Z_2) > r);$$

r è un valore fissato che definisce la quantità teorica di malati, in questo caso fissata al 25%

Con i valori del test e delle covariate, viene calcolata di volta in volta la probabilità per di un soggetto di essere verificato (secondo, poiché si ignora il valore di D, le assunzioni MAR) utilizzando un opportuno modello.

$$\pi = j(T, X)$$

Calcolata questa probabilità attraverso una variabile casuale di Bernuolli di parametro π ad ogni osservazione viene assegnato il valore di V. Per ogni replicazione tutti i valori a parte r vengono ricalcolati.

Capitolo 4

Risultati

Di seguito si riportano i grafici e le tabelle che sintetizzano i risultati delle varie simulazioni. Nei grafici e nelle colonne numeriche di destra delle tabelle vengono riportate le medie delle stime dell'AUC calcolate via Jackknife, via Monte Carlo e la full, cioè quella calcolata presupponendo che tutte le osservazioni siano verificate. A sinistra invece si hanno il valore della varianza dello stimatore dell'AUC calcolato via Monte Carlo e la media delle varianze Jackknife di $\widehat{AUC}_{KNN}$; per quest'ultime, solo nella tabella è riportata pure la mediana.

4.1 Scenario A

Per lo scenario A sono stati utilizzati i seguenti dati:

$A = 5000$ numero di replicazioni Monte Carlo

$n = \{200, 400\}$ numerosità del campione

$k = \{1, 3\}$ numero di vicini-più-vicini

$f(Z_1, Z_2) = g(Z_1, Z_2) = Z_1 + Z_2$

$\alpha = \{0.5, 1, 1.5\}$

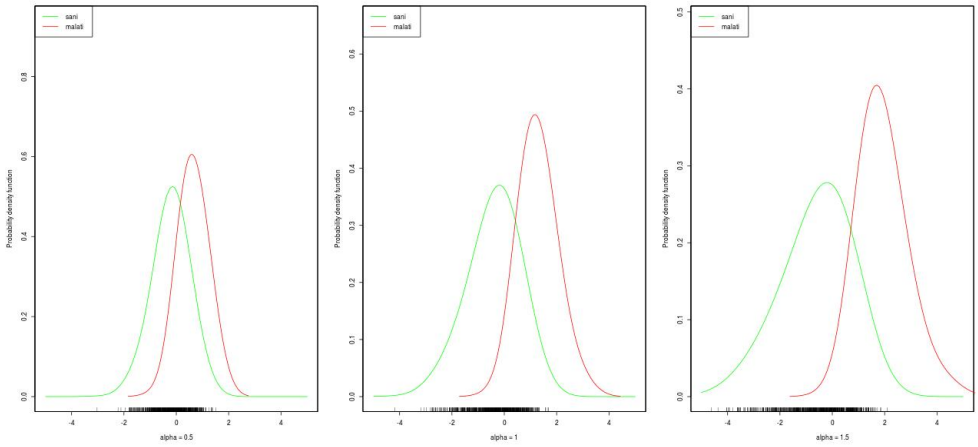
$h(Z_1, Z_2) = \alpha(Z_1 + Z_2)$

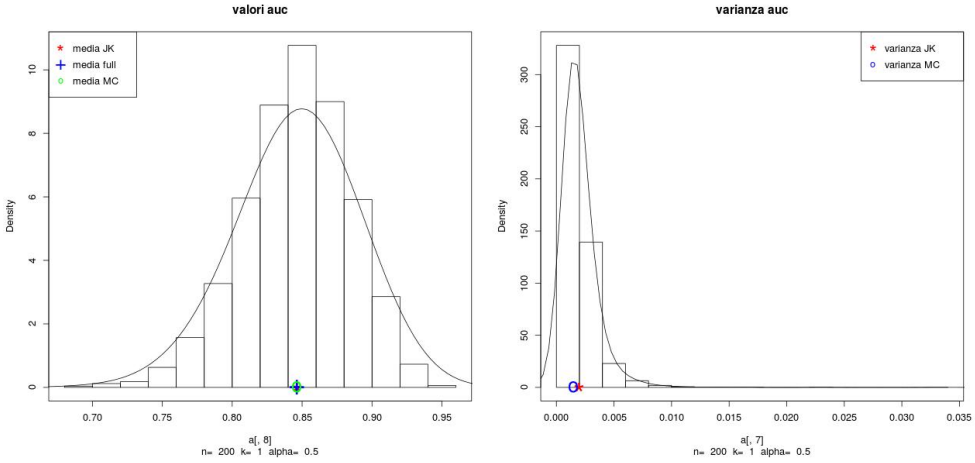
$r = z_{0.75} = 0.6744898$

$j(T, X) = \frac{e^q}{1+e^q}$ dove $q = 0.05 + 0.9T + 0.7X$

Con questa scelta di $j(T, X)$ si ottiene che la probabilità π media di un soggetto di essere verificato si aggira a intorno a 0.51.

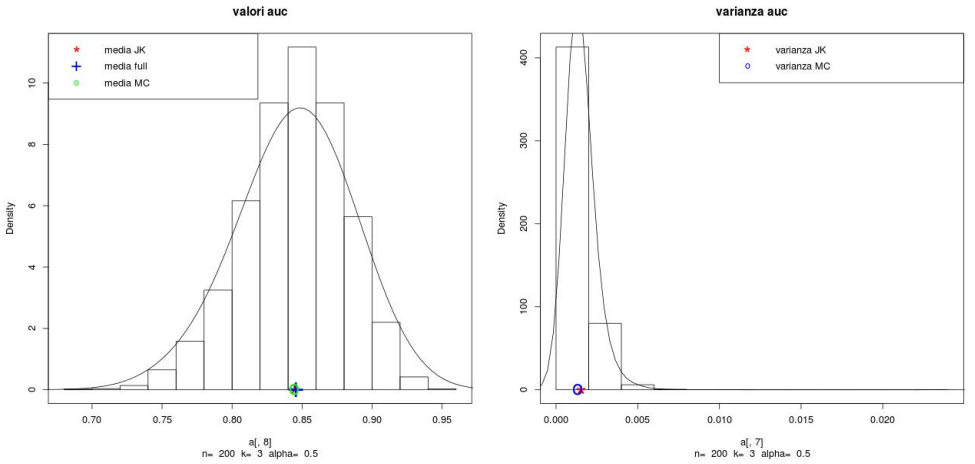
Le due numerosità campionarie servono a rendere un'idea della capacità di stima del metodo Jackknife, mentre, come mostrato dai grafici sottostanti, l'aumento di α aumenta il livello di separazione tra i due gruppi, "sani" e "malati".





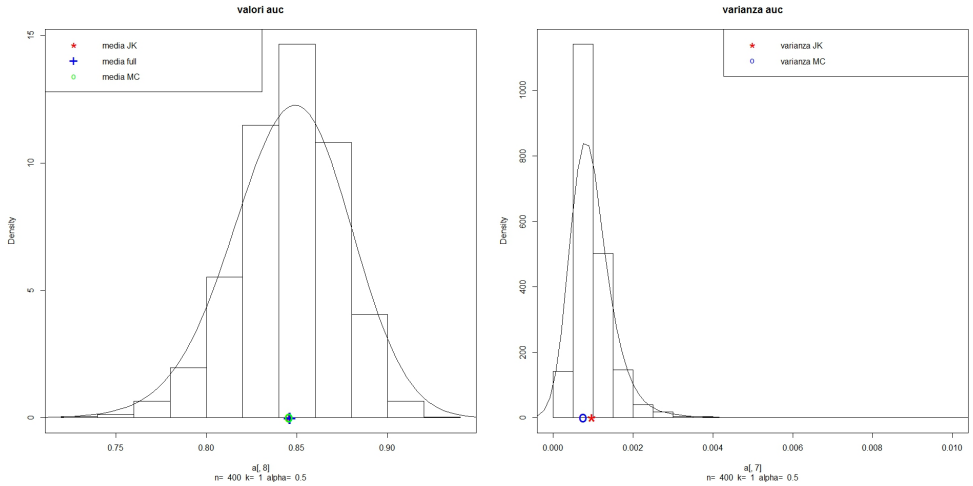
media Monte Carlo	0.8461404	varianza Monte Carlo	0.001479204
media jackknife	0.8461394	varianza jackknife	0.0019867
media full	0.8461532	mediana varianza JK	0.001620021

Tab. 4.1: n=200, k=1, alpha=0,5



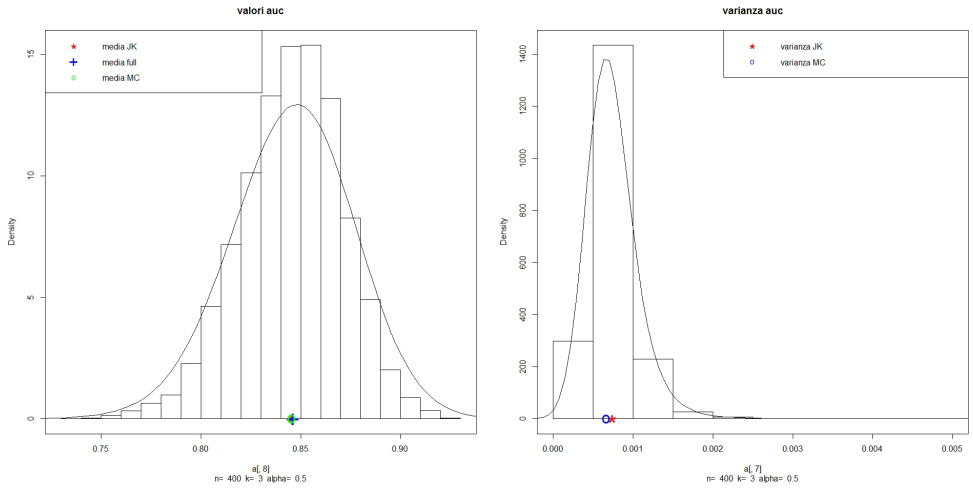
media Monte Carlo	0.8445541	varianza Monte Carlo	0.001325675
media jackknife	0.8445485	varianza jackknife	0.001511717
media full	0.8455736	mediana varianza JK	0.001347903

Tab. 4.2: n=200, k=3, alpha=0,5



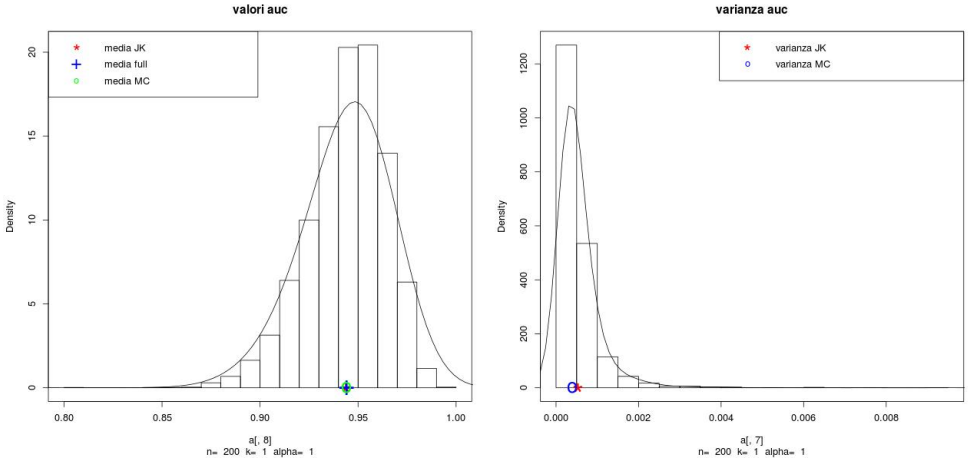
media Monte Carlo	0.8456008	varianza Monte Carlo	0.0007484141
media jackknife	0.845601	varianza jackknife	0.0009735334
media full	0.8460063	mediana varianza JK	0.0008576286

Tab. 4.3: n=400, k=1, alpha=0,5



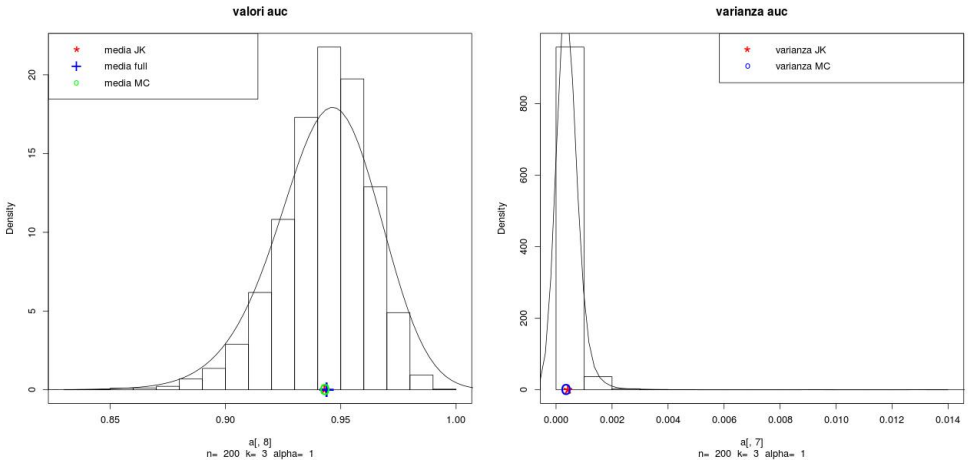
media Monte Carlo	0.8455352	varianza Monte Carlo	0.0006663436
media jackknife	0.8455337	varianza jackknife	0.0007405751
media full	0.8460728	mediana varianza JK	0.0006893221

Tab. 4.4: n=400, k=3, alpha=0,5



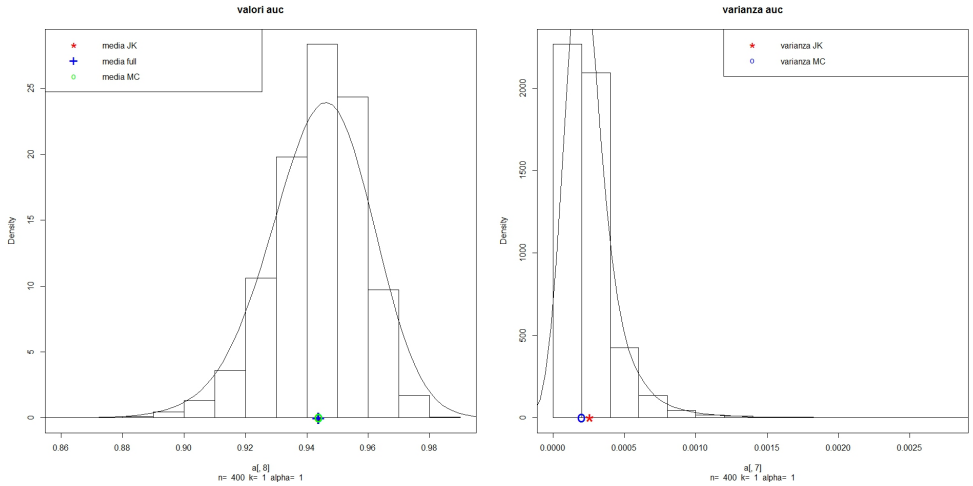
media Monte Carlo	0.9439049	varianza Monte Carlo	0.0003984796
media jackknife	0.9439048	varianza jackknife	0.0005342412
media full	0.9439056	mediana varianza JK	0.0003990252

Tab. 4.5: n=200, k=1, alpha=1



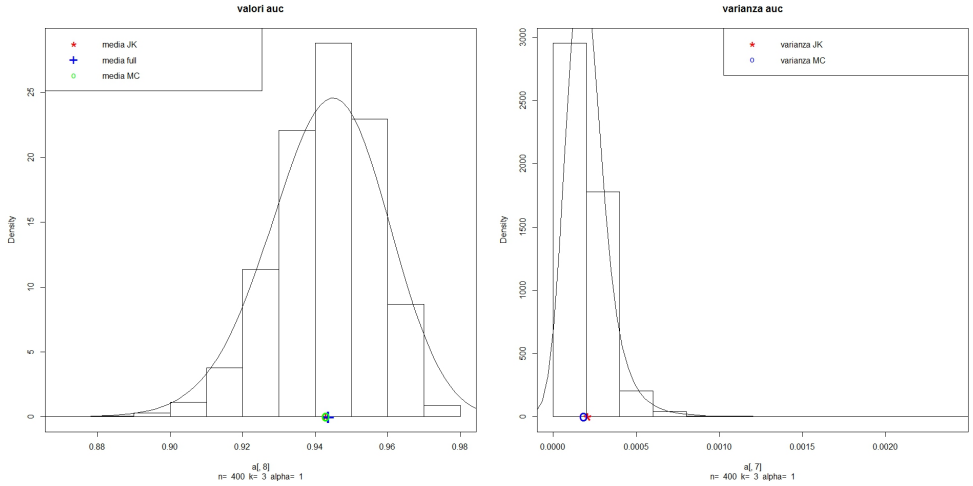
media Monte Carlo	0.9430624	varianza Monte Carlo	0.000370723
media jackknife	0.9430579	varianza jackknife	0.0004252161
media full	0.9436628	mediana varianza JK	0.0003477284

Tab. 4.6: n=200, k=3, alpha=1



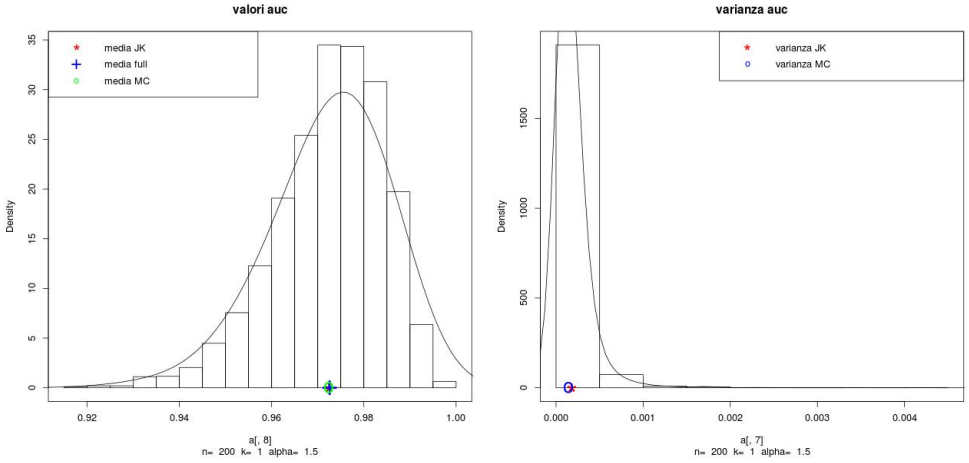
media Monte Carlo	0.9438799	varianza Monte Carlo	0.0001989812
media jackknife	0.9438797	varianza jackknife	0.0002574874
media full	0.943782	mediana varianza JK	0.000210718

Tab. 4.7: n=400, k=1, alpha=1



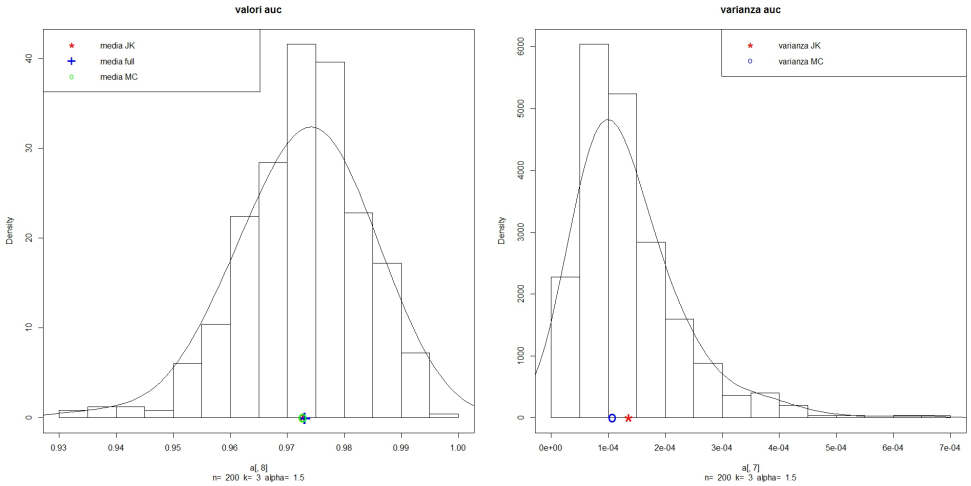
media Monte Carlo	0.9430401	varianza Monte Carlo	0.0001823016
media jackknife	0.9430396	varianza jackknife	0.0002071397
media full	0.9435727	mediana varianza JK	0.0001802563

Tab. 4.8: n=400, k=3, alpha=1



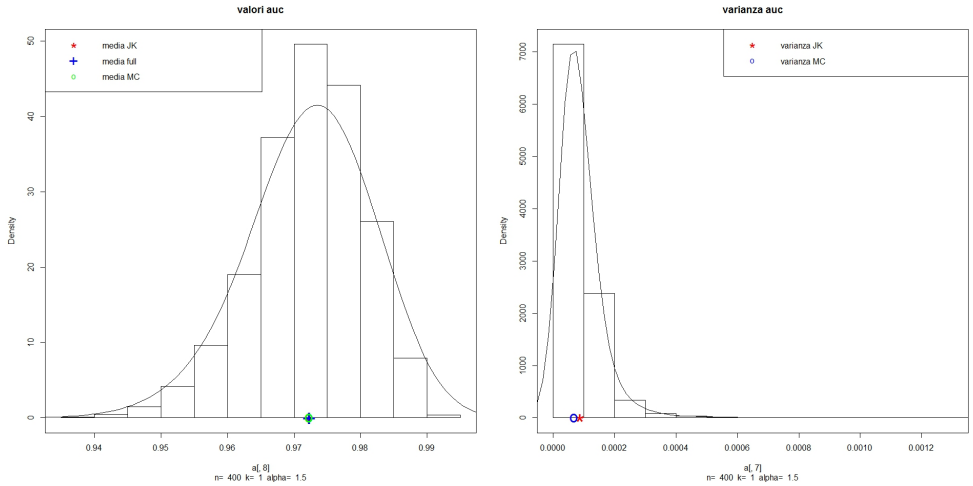
media Monte Carlo	0.9723861	varianza Monte Carlo	0.0001425642
media jackknife	0.9723861	varianza jackknife	0.0001811991
media full	0.9724941	mediana varianza JK	0.0001304193

Tab. 4.9: n=200, k=1, alpha=1,5



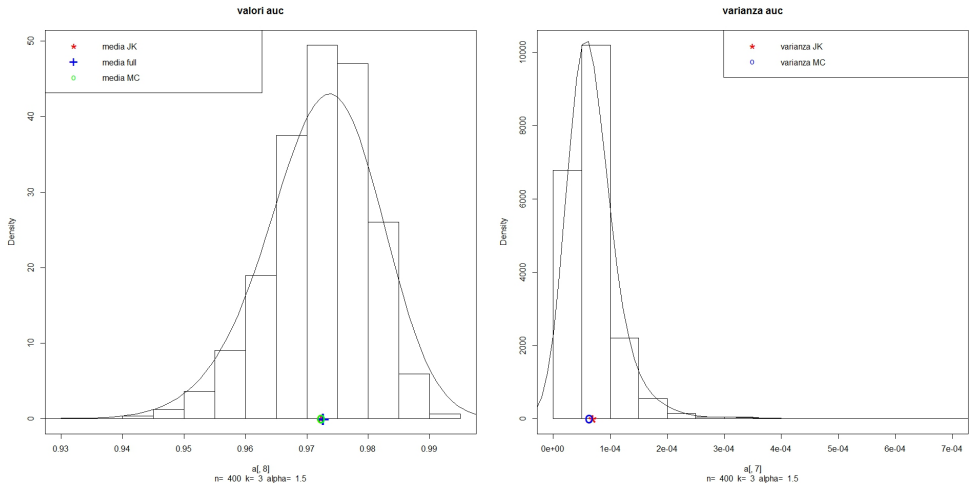
media Monte Carlo	0.9727751	varianza Monte Carlo	0.0001079379
media jackknife	0.9727728	varianza jackknife	0.0001359186
media full	0.9730367	mediana varianza JK	0.0001149666

Tab. 4.10: n=200, k=3, alpha=1,5



media Monte Carlo	0.9722783	varianza Monte Carlo	0.00006761377
media jackknife	0.972278	varianza jackknife	0.0000876891
media full	0.9722939	mediana varianza JK	0.00007170353

Tab. 4.11: n=400, k=1, alpha=1,5



media Monte Carlo	0.9723821	varianza Monte Carlo	0.00006323777
media jackknife	0.9723814	varianza jackknife	0.00006982942
media full	0.9726185	mediana varianza JK	0.00006084196

Tab. 4.12: n=400, k=3, alpha=1,5

4.2 Scenario B

Per lo scenario B sono stati utilizzati i seguenti dati:

$A = 5000$; numero di replicazioni Monte Carlo

$n = \{200, 400\}$; numerosità del campione

$k = \{1, 3\}$; numero di vicini-più-vicini

$f(Z_1, Z_2) = \sqrt{2}(Z_1^2 + Z_2^2)$

$g(Z_1, Z_2) = \exp\,[2(Z_1 Z_2)^2]$

$h(Z_1, Z_2) = 2(Z_1 Z_2)^2$

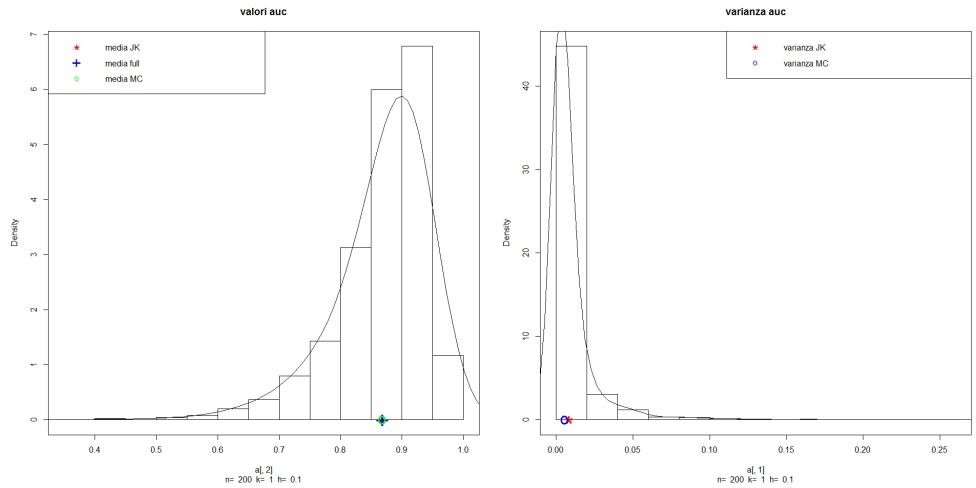
$r = 1.4516$; 75-esimo percentile della distribuzione di g

$\pi(T, X) = 0.05 + \delta I[T > 1.2] + (0.95 - \delta) I[X > 1.95]$

$\delta = \{0.1, 0.5, 0.9\}$

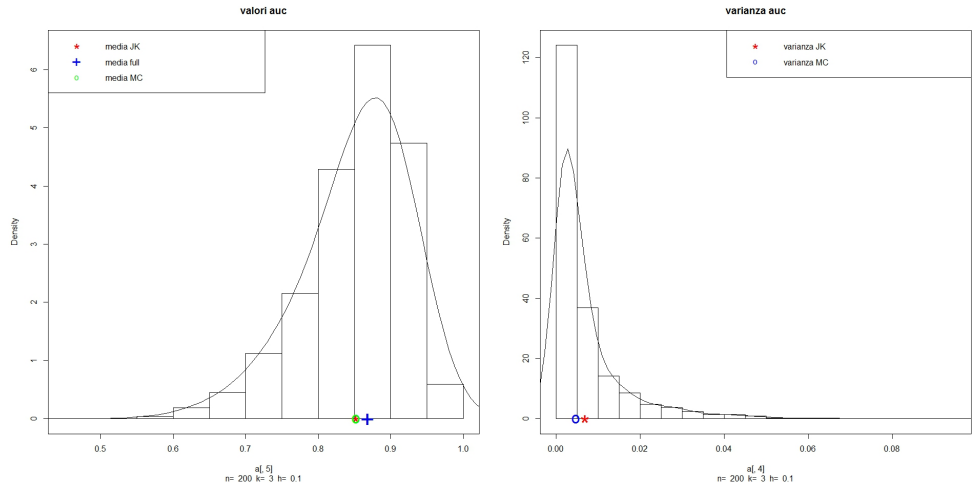
Al variare di δ cambia la probabilità di un soggetto di essere verificato. Questa probabiltà vale:

1	se $T > 1.2$ e $X > 1.95$;
$1 - \delta$	se $T \leq 1.2$ e $X > 1.95$;
$0.05 + \delta$	se $T > 1.2$ e $X \leq 1.95$;
0.05	altrimenti.



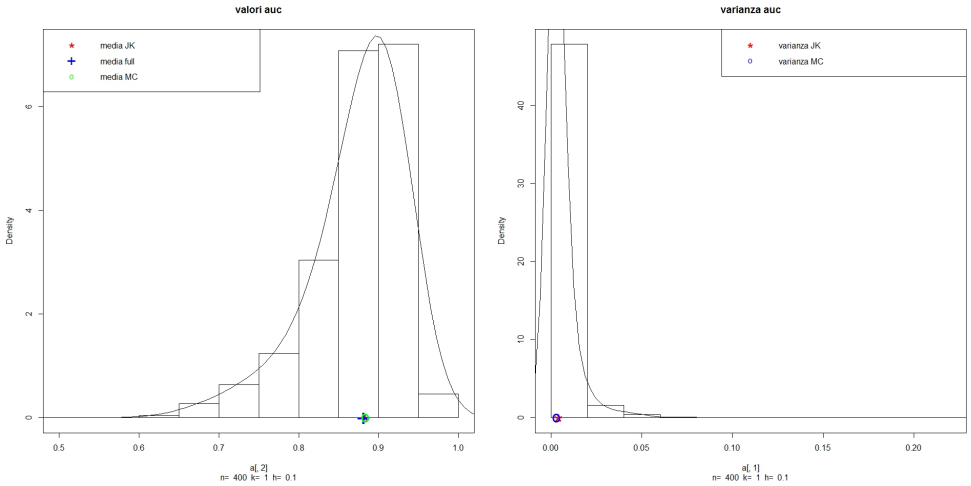
media Monte Carlo	0.8680733	varianza Monte Carlo	0.005511302
media jackknife	0.8680623	varianza jackknife	0.008697657
media full	0.8680099	mediana varianza JK	0.003096443

Tab. 4.13: n=200, k=1, h=0,1



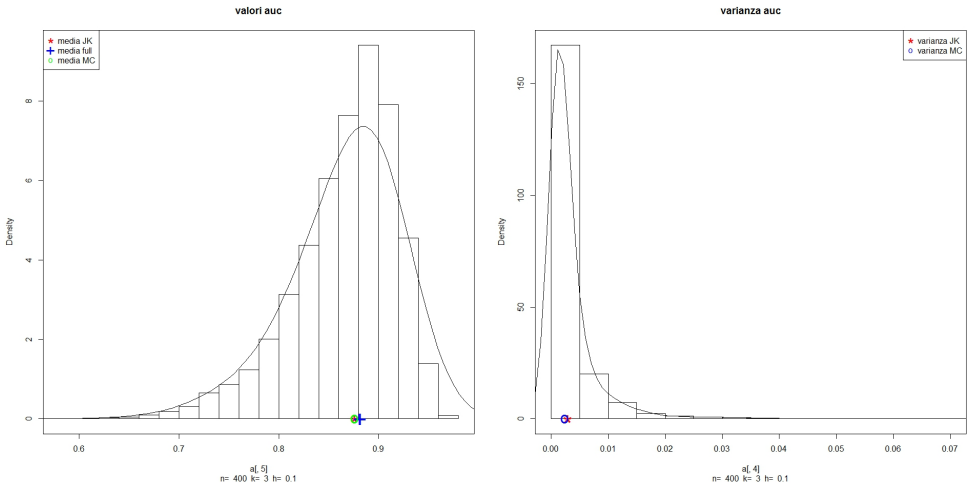
media Monte Carlo	0.8524464	varianza Monte Carlo	0.004751865
media jackknife	0.8523345	varianza jackknife	0.006911563
media full	0.8679728	mediana varianza JK	0.003524206

Tab. 4.14: n=200, k=3, h=0,1



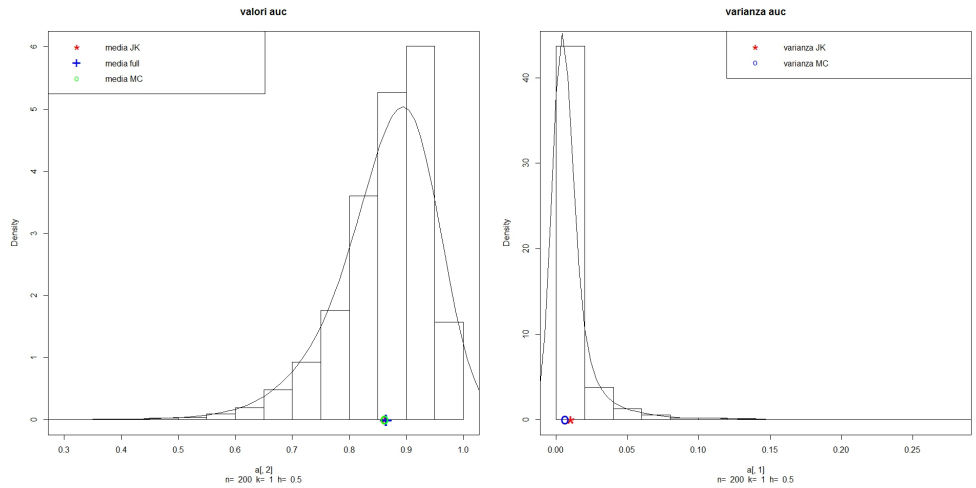
media Monte Carlo	0.8835629	varianza Monte Carlo	0.00297939
media jackknife	0.8835603	varianza jackknife	0.004120606
media full	0.8814374	mediana varianza JK	0.001849162

Tab. 4.15: n=400, k=1, h=0,1



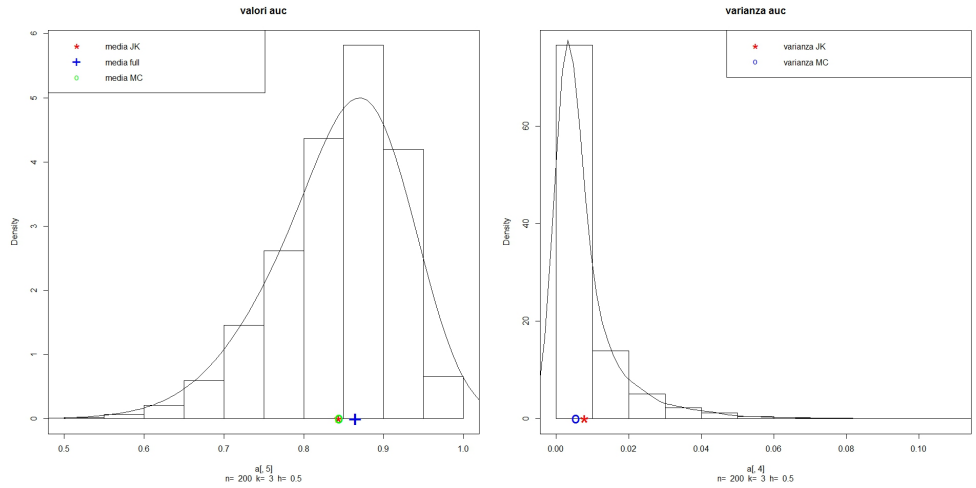
media Monte Carlo	0.8763327	varianza Monte Carlo	0.002367634
media jackknife	0.8763178	varianza jackknife	0.002932154
media full	0.8814376	mediana varianza JK	0.001638951

Tab. 4.16: n=400, k=3, h=0,1



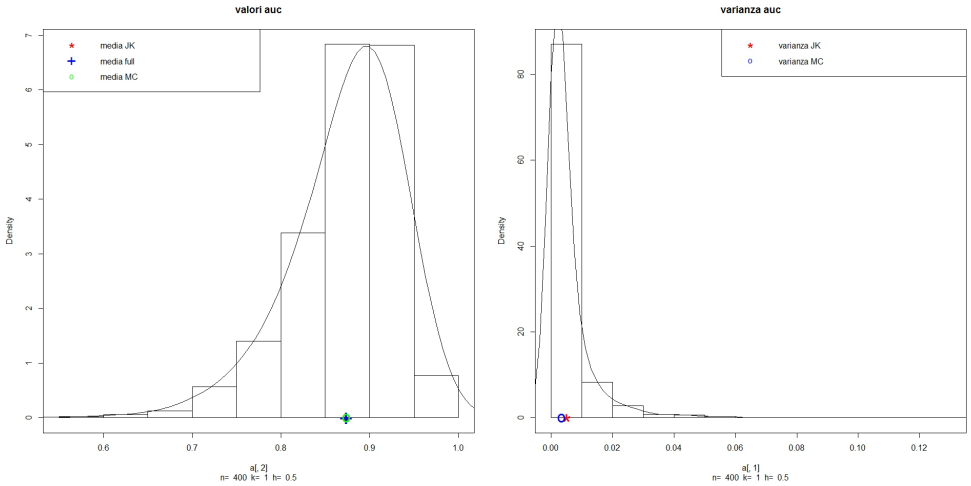
media Monte Carlo	0.8625465	varianza Monte Carlo	0.006268507
media jackknife	0.8625287	varianza jackknife	0.01032227
media full	0.8643488	mediana varianza JK	0.00444464

Tab. 4.17: n=200, k=1, h=0,5



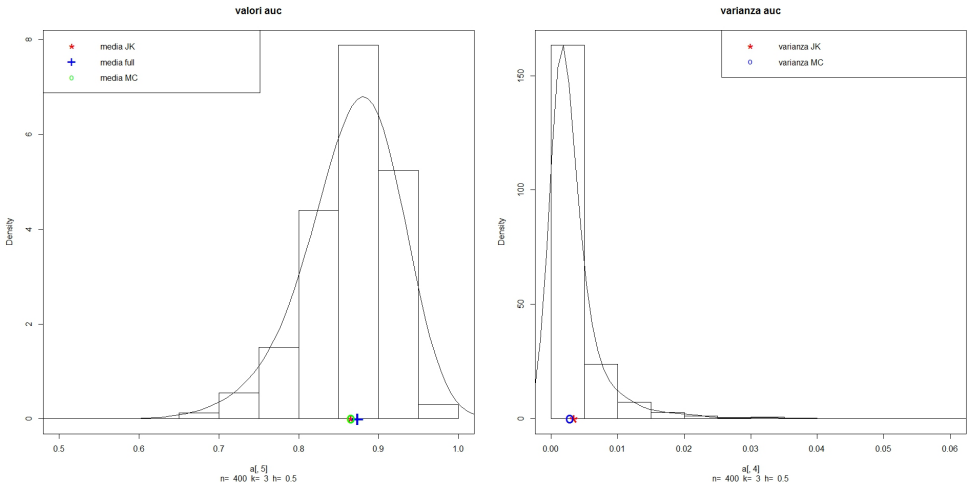
media Monte Carlo	0.8443381	varianza Monte Carlo	0.005401287
media jackknife	0.8442121	varianza jackknife	0.007877943
media full	0.8642772	mediana varianza JK	0.00428943

Tab. 4.18: n=200, k=3, h=0,5



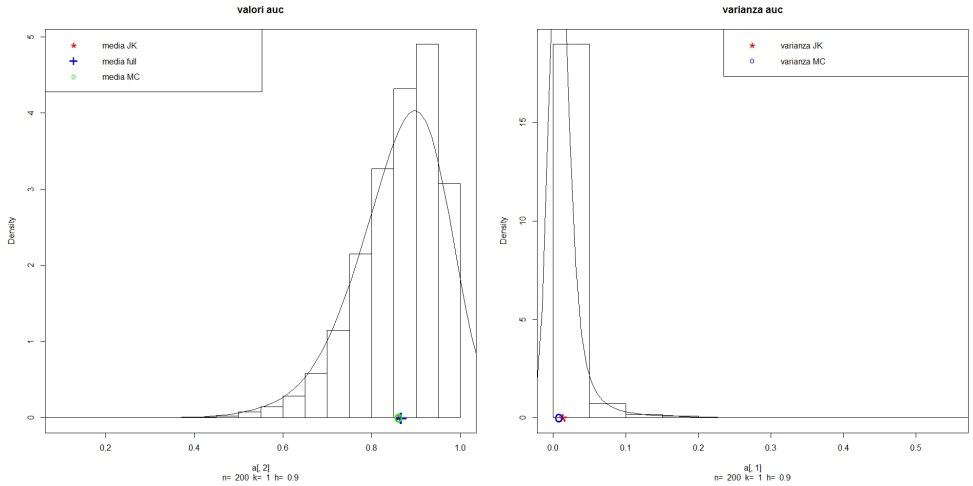
media Monte Carlo	0.874162	varianza Monte Carlo	0.003412597
media jackknife	0.8741591	varianza jackknife	0.005076814
media full	0.8735351	mediana varianza JK	0.002391384

Tab. 4.19: n=400, k=1, h=0,5



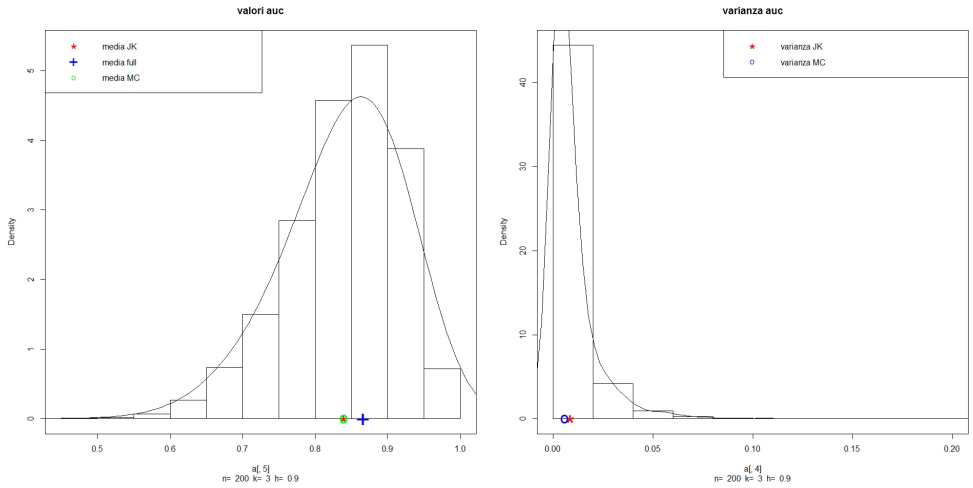
media Monte Carlo	0.8655761	varianza Monte Carlo	0.002790737
media jackknife	0.865554	varianza jackknife	0.003424871
media full	0.8736359	mediana varianza JK	0.00201403

Tab. 4.20: n=400, k=3, h=0,5



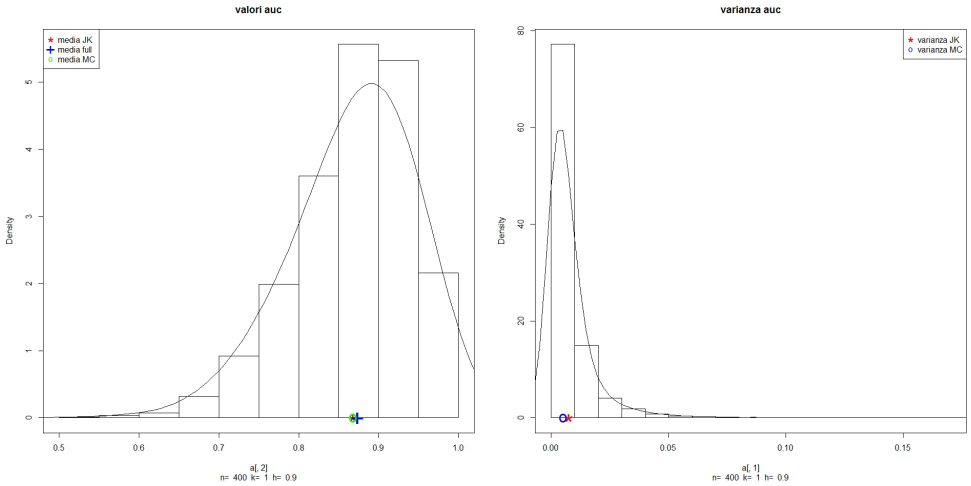
media Monte Carlo	0.8613545	varianza Monte Carlo	0.008400967
media jackknife	0.8613387	varianza jackknife	0.01363321
media full	0.8655308	mediana varianza JK	0.005992634

Tab. 4.21: n=200, k=1, h=0,9



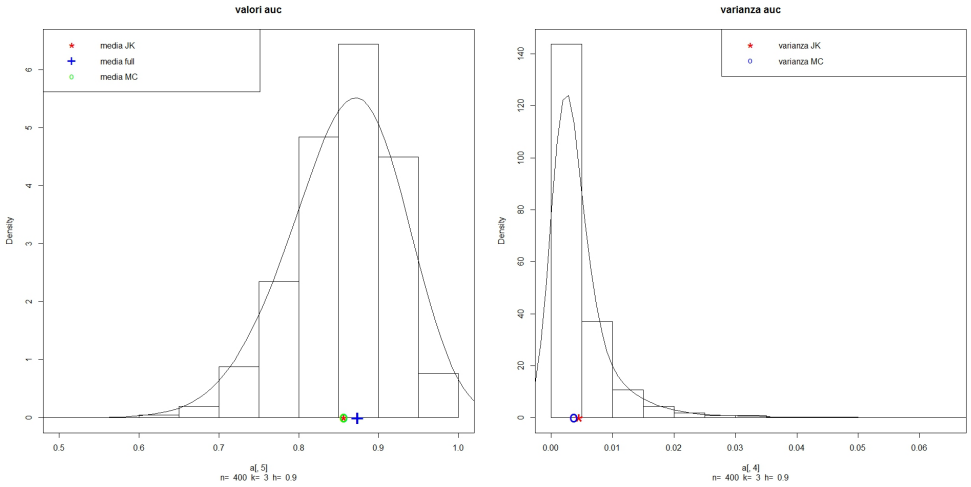
media Monte Carlo	0.8395441	varianza Monte Carlo	0.005847114
media jackknife	0.839451	varianza jackknife	0.008756539
media full	0.8654877	mediana varianza JK	0.004499446

Tab. 4.22: n=200, k=3, h=0,9



media Monte Carlo	0.8682113	varianza Monte Carlo	0.005056724
media jackknife	0.8682114	varianza jackknife	0.007553249
media full	0.8732038	mediana varianza JK	0.004269387

Tab. 4.23: n=400, k=1, h=0,9



media Monte Carlo	0.8569479	varianza Monte Carlo	0.00372209
media jackknife	0.8569249	varianza jackknife	0.004596463
media full	0.8731879	mediana varianza JK	0.003002504

Tab. 4.24: n=400, k=3, h=0,9

Capitolo 5

Un esempio su dati reali

Viene ora analizzato un data-set di dati reali che riguardano le diagnosi per il cancro al seno. Le pazienti sono 569, di cui 327 verificate. Tra le verificate sono stati diagnosticati 118 cancri benigni e 209 maligni.

Il data-set contiene i seguenti dati:

- T: il valore del test ($\mu = 16.27, \sigma^2 = 23.26$);
- X: il valore di una covariata($\mu = 0.115, \sigma^2 = 0.004$);
- V: vale 1 se il è verificato, 0 altrimenti;
- Disease: in cui se l'osservazione è verificata è riportato se il cancro e maligno o meno.

Si calcola ora il valore dell'$\widehat{AUC}_{KNN}$ e se ne stima la varianza tramite il metodo Jackknife per i valori di k 1 e 3 utilizzando quattro distanze diverse: la distanza Euclidea $d(x_i, x_j) = \sqrt{\sum_{h=1}^{2}(x_{ih} - x_{ih})^2}$, la distanza di Lagrange $d(x_i, x_j) = \max_{h\in\{1,2\}} |x_{ih} - x_{ih}|$, la distanza di Mahalonobis $d(x_i, x_j) = \sqrt{(x_i - x_j)^T \hat{\Sigma}^{-1}(x_i - x_j)}$ e la distanza di Manhattan $d(x_i, x_j) = \sum_{h=1}^{2} |x_{ih} - x_{ih}|$. Viene poi calcolato un intervallo di confidenza al 95% e al 99% per l'$\widehat{AUC}_{KNN}$ basandosi sulle assunzione di normalità di questo stimatore. Si riportano di seguito i valori di k, dell'$\widehat{AUC}_{KNN}$ e di quelli stimati col Jackknife, viene poi riportata la stima della varianza

dello stimare via Jackknife $\hat{V}_{JK}$ e i due intervalli di confidenza calcolati come $\widehat{AUC}_{KNN} \pm 1.96\sqrt{\hat{V}_{JK}}$ e $\widehat{AUC}_{KNN} \pm 2.56\sqrt{\hat{V}_{JK}}$.

k	$\widehat{AUC}_{KNN}$	$\widehat{AUC}_{JK}$	Varianza JK	IC 95%	IC 99%
1	0.9738	0.9738	3.515785e-05	(0.9622, 0.9854)	(0.9586, 0.989)
3	0.9578	0.9578	4.824759e-05	(0.9442, 0.9714)	(0.94, 0.9756)

Tab. 5.1: Risultati dell'analisi dei dati reali con la distanza Euclidea

k	$\widehat{AUC}_{KNN}$	$\widehat{AUC}_{JK}$	Varianza JK	IC 95%	IC 99%
1	0.9738	0.9738	8.641671e-04	(0.9162,1)	(0.8981,1)
3	0.9583	0.9583	4.991311e-04	(0.9146,1)	(0.9008,1)

Tab. 5.2: Risultati dell'analisi dei dati reali con la distanza di Lagrange

k	$\widehat{AUC}_{KNN}$	$\widehat{AUC}_{JK}$	Varianza JK	IC 95%	IC 99%
1	0.9506	0.9506	7.466465e-04	(0.8970,1)	(0.8801,1)
3	0.9612	0.9613	2.170532e-04	(0.9324, 0.9901)	(0.9232,0.9991)

Tab. 5.3: Risultati dell'analisi dei dati reali con la distanza di Mahalanobis

k	$\widehat{AUC}_{KNN}$	$\widehat{AUC}_{JK}$	Varianza JK	IC 95%	IC 99%
1	0.9738	0.9738	10.33851e-04	(0.9108, 1)	(0.8910, 1)
3	0.95712	0.9572	3.133556e-04	(0.9224, 0.9919)	(0.9116, 1)

Tab. 5.4: Risultati dell'analisi dei dati reali con la distanza di Manhattan

Capitolo 6

Conclusione

Entrambi gli scenari mostrano come il metodo Jackknife possa essere utilizzato per la stima della varianza e che con un K e una numerosità maggiore il risultato è migliore. Questo risultato ha permesso di poter utilizzare l'approccio per lo studio di dati reali. I grafici inoltre sembrano confermare la distribuzione normale dello stimatore $\widehat{AUC}_{KNN}$, data la forma dell'istogramma delle medie che è a "campana". Il tempo di elaborazione per ogni set diverso di dati è stato di circa 3 ore per $n = 200$ e 18 per $n = 400$ per entambi gli scenari.

Appendice A

Codice Utilizzato

A.1 Scenario A

```
n <- 200 # numerosità campionaria: 200 400
A<- 5000
k <- 3# 1, 3
c<-0.2 # 0.2 0.5 0.8
alpha <- 1.5 #  0.5, 1, 1.5
w <- 0.05
wt <- 0.9
wx <- 0.7
a<-matrix(0,A,12)
r <- qnorm(0.75)
for(i in 1:A)
    {
    oo<-A-i
    cat(oo,' ')
    if(i/10==floor(i/10)) cat('\n')
    z1 <- rnorm(n, sd=sqrt(0.5))
    z2 <- rnorm(n, sd=sqrt(0.5))
    e1 <- rnorm(n, sd=sqrt(0.25))
    e2 <- rnorm(n, sd=sqrt(0.25))
    zz<-z1+z2
    D <- as.integer(zz>r)
    T <- (z1+z2) * alpha + e1
```

```
23      X <- z1 + z2 + e2
24      y <- w + wt * T + wx * X
25      p <- exp(y)/(1+exp(y))
26      V <- rbinom(n, 1, p)
27      T1<-T[V==1]
28      T0<-T[V==0]
29      X1<-X[V==1]
30      X0<-X[V==0]
31      D1<-D[V==1]
32      D0<-D[V==0]
33      V1<-V[V==1]
34      V0<-V[V==0]
35      l1<-sum(V==1)
36      D<-c(D1,D0)
37      X<-c(X1,X0)
38      T<-c(T1,T0)
39      V<-c(V1,V0)
40      rocfull<-tau.hat(T,D)
41      sefull<-se.full(T,D,c)
42      spfull<-sp.full(T,D,c)
43      dist<-distanza(cbind(X1,T1),cbind(X0,T0))
44      a[i,]<-c(jackknife(T,X,D,V,k,c,l1,dist),rocfull,sefull,spfull)
45  }
46  sozz<-apply(a,2,mean)
47  titoli <- c('n'= n,'k'= k,'c' =c,'alpha' = alpha)
48  titoli
49  full<-c('se.full'=sozz[11], 'var.se.full'=var(a[,11]),
        'sp.full'=sozz[12], 'var.sp.full'=var(a[,12]),
        'roc.full'=sozz[10], 'var.roc.full'=var(a[,10]))
50  soluzse<-c('val.jack.se'=sozz[2], 'var.jack.se'=sozz[1],
        'val.mont.se'=sozz[3], 'var.mont.se'=var(a[,3]))
51  soluzsp<-c('val.jack.sp'=sozz[5], 'var.jack.sp'=sozz[4],
        'val.mont.sp'=sozz[6], 'var.mont.sp'=var(a[,6]))
52  soluzauc<-c('val.jack.auc'=sozz[8],
        'var.jack.auc'=sozz[7], 'val.mont.auc'=sozz[9],
        'var.mont.auc'=var(a[,9]))
53  full
```

```
54 soluzse
55 soluzsp
56 soluzauc
```

Da riga 1 a 26 vengono generati i dati, a è la matrice che registra nella sua i-esima riga i risultati dell'i-esima replicazione;
da riga 27 a 39 vengono ordinati i dati in modo che siano divisi fra verificati e no;
Da riga 40 a 42 vengono calcolate l'AUC, la specificità e la sensitività presupponendo che i dati siano verificati;
Alla riga 43 vengono calcolate le distanze euclidee fra i dati verificati e quelli no;
Alla riga 44 vengono calcolati i risultati della singola replicazione;
Alla riga 46 si fa la media dei valori calcolate nelle singole replicazioni;
Dalla riga 49 alla riga 56 vengono ordinati e riportati i risultati.

A.2 Scenario B

```
1  n <- 400 # numerosità campionaria 200 , 400
2  A<- 5000
3  k <- 1# 1,  3
4  h <- 0.1 #  0.1 0.5 0.9
5  c<-0.2 # 0.2 0.5 0.8
6  a <-matrix(0,A,12)
7  r <- quantile(exp(2*(rnorm(1001,
       sd=sqrt(0.5))*rnorm(1001, sd=sqrt(0.5)))^2),0.75)
8  for(i in 1:A)
9      {
10     oo<-A-i
11     cat(oo,' ')
12     if(i/10==floor(i/10)) cat('\n')
13     z1 <- rnorm(n, sd=sqrt(0.5))
14     z2 <- rnorm(n, sd=sqrt(0.5))
15     e1 <- rnorm(n, sd=sqrt(0.25))
16     e2 <- rnorm(n, sd=sqrt(0.25))
17     zz<-exp(2*(z1*z2)^2)
```

```
18      D <- as.integer(zz>r)
19      T <- 2*(z1*z2)^2  + e1
20      X <- sqrt(2)*(z1^2 + z2^2) + e2
21      p <-
            0.05+h*as.integer(T>1.2)+(0.95-h)*as.integer(X>1.95)
22      V <- rbinom(n, 1, p)
23      T1<-T[V==1]
24      T0<-T[V==0]
25      X1<-X[V==1]
26      X0<-X[V==0]
27      D1<-D[V==1]
28      D0<-D[V==0]
29      V1<-V[V==1]
30      V0<-V[V==0]
31      l1<-sum(V==1)
32      D<-c(D1,D0)
33      X<-c(X1,X0)
34      T<-c(T1,T0)
35      V<-c(V1,V0)
36      rocfull<-tau.hat(T,D)
37      sefull<-se.full(T,D,c)
38      spfull<-sp.full(T,D,c)
39      dist<-distanza(cbind(X1,T1),cbind(X0,T0))
40      a[i,]<-c(jackknife(T,X,D,V,k,c,l1,dist),rocfull,sefull,spfull)
41  }
42  sozz<-apply(a,2,mean)
43  titoli <- c('n'= n,'k'= k,'c' =c,'alpha' = alpha)
44  titoli
45  full<-c('se.full'=sozz[11], 'var.se.full'=var(a[,11]),
        'sp.full'=sozz[12], 'var.sp.full'=var(a[,12]),
        'roc.full'=sozz[10], 'var.roc.full'=var(a[,10]))
46  soluzse<-c('val.jack.se'=sozz[2], 'var.jack.se'=sozz[1],
        'val.mont.se'=sozz[3], 'var.mont.se'=var(a[,3]))
47  soluzsp<-c('val.jack.sp'=sozz[5], 'var.jack.sp'=sozz[4],
        'val.mont.sp'=sozz[6], 'var.mont.sp'=var(a[,6]))
48  soluzauc<-c('val.jack.auc'=sozz[8],
        'var.jack.auc'=sozz[7], 'val.mont.auc'=sozz[9],
```

```
    'var.mont.auc'=var(a[,9]))
full
soluzse
soluzsp
soluzauc
```

Per la spiegazione del codice vedere quella del scenario A

A.3 Funzioni

```
jackknife<- function(T,X,D,V,k,c,l1,dist)
{

    n<-length(T)
    phatse<-double(n)
    phatsp<-double(n)
    phatauc<-double(n)
    for(i in 1:l1)
    {
        dj<-D[-i]
        tj<-T[-i]
        xj<-X[-i]
        vj<-V[-i]
        distj<-dist[-i,]
        #tolgo l'i-esima osservazione
        rhoj<-rho.knn(distj,k,dj,(n-l1))
        rhot<-c(rep(1,l1-1),rhoj)
        #calcolo il rho per il nuovo campione
        D.tildej<-vj*dj+(1-vj)*rhot
        phatauc[i]<-tau.hat(tj,D.tildej)
        #calcolo l'auc per l'i-esimo campione jackknife
        phatse[i]<-se.hat(tj,dj,vj,c,rhot)
        #calcolo la sensitività per l'i-esimo campione
            jackknife
        phatsp[i]<-sp.hat(tj,dj,vj,c,rhot)
        #calcolo la specificità per l'i-esimo campione
            jackknife
    }
```

```
27      for(i in (l1+1):n)
28      {
29          dj<-D[-i]
30          tj<-T[-i]
31          xj<-X[-i]
32          vj<-V[-i]
33          distj<-dist[,-(i-l1)]
34          #tolgo l'i-esima osservazione
35          rhoj<-rho.knn(distj,k,dj,(n-l1-1))
36          #calcolo il rho per il nuovo campione
37          rhot<-c(rep(1,l1),rhoj)
38          D.tildej<-vj*dj+(1-vj)*rhot
39          phatauc[i]<-tau.hat(tj,D.tildej)
40          #calcolo l'auc per l'i-esimo campione jackknife
41          phatse[i]<-se.hat(tj,dj,vj,c,rhot)
42          #calcolo la sensitività per l'i-esimo campione
                jackknife
43          phatsp[i]<-sp.hat(tj,dj,vj,c,rhot)
44          #calcolo la specificità per l'i-esimo campione
                jackknife
45      }
46      rho<-rho.knn(dist,k,D,(n-l1))
47      rhow<-c(rep(1,l1),rho)
48      #calcolo il rho per il campione con tutte le
            osservazioni
49      D.tilde<-V*D+(1-V)*rhow
50      pauc<-tau.hat(T,D.tilde)
51      #calcolo l'auc su tutte le osservazioni
52      pse<-se.hat(T,D,V,c,rhow)
53      #calcolo la sensitività su tutte le osservazioni
54      psp<-sp.hat(T,D,V,c,rhow)
55      #calcolo la specificità su tutte le osservazioni
56      pjackse<-(mean(phatse))
57      #calcolo la stima della sensitività con jackknife
58      pjacksp<-(mean(phatsp))
59      #calcolo la stima della specificità con jackknife
60      ppseudose<-n*pse-(n-1)*phatse
```

```
61      #calcolo i pseudo-valori jackknife per la sensitività
62      ppseudosp<-n*psp-(n-1)*phatsp
63      #calcolo i pseudo-valori jackknife per la specificità
64      vse<-var(ppseudose)/n
65      #calcolo la varianza jackknife per la sensitività
66      vsp<-var(ppseudosp)/n
67      #calcolo la varianza jackknife per la specificità
68      pjackauc<-mean(phatauc)
69      #calcolo la stima dell'auc con jackknife
70      ppseudoauc<-n*pauc-(n-1)*phatauc
71      #calcolo i pseudo-valori jackknife per l'auc
72      vauc<-var(ppseudoauc)/n
73      #calcolo la varianza jackknife per l'auc
74      sol<-c(vse,pjackse,pse,vsp,pjacksp,psp,vauc,pjackauc,pauc)
75      sol
76  }
```

Questa funzione calcola tramite il procedimento illustrato nel paragrafo 2.2 i valori della varianza Jackknife, V_{JK}, della stima Jackknife, $\bar{\hat{\theta}}_{(\bullet)}$, e della stima su tuttele osservazioni $\hat{\theta}$

```
1  tau.hat <- function(T, D)
2  {
3    n <- length(T)
4    tau <- 0
5    num <- 0
6    den <- 0
7    for(i in 1:n)
8    {
9      t<-T[-i]
10     ti<-T[i]
11          num <- num + sum((t>ti) * D[-i] * (1-D[i]))
12          den <- den + sum(D[-i] * (1-D[i]))
13   }
14   tau <- num / den
15   tau
16 }
```

Questa funzione corrisponde a (1.3) e calcola il valore dell'AUC dato il valore di T e di D, nel nostro caso la funzione jackknife passa il valore di $\tilde{D} = \{V_i D_i + (1 - V_i)\hat{\rho}_{Ki}\}$.

Nel ciclo for ogni volta viene confrontato il vettore T privato di una osservazione con l'osservazione stessa.

```
rho.knn <- function(mat,k,D,l1)
    {
    id <-colmink(mat,k)
    #trova le posizioni delle k minime distanze
    rho <- double(l1)
    for(i in 1:l1)
        {
        rho[i] <- mean(D[id[,i]])
        #calola i vari rho per le varie osservazioni
        }
    rho
    }
```

Questa funzione calcola il valore della media di D dei KNN (1.4)

```
colmink <- function(m,k)
    {
    mini<-matrix(apply(m,2,min.k,k),k)
    mini
    }
```

Questa funzione trova la posizione dei k minimi di ogni colonna di una matrice

```
min.k<-function(x,k)

    {
    id<-double(k)
    for(i in 1:k)
        {
        id[i] <- match(min(x), x)
        x[id[i]]<- Inf
        }
    id
```

```
11      }
```

Questa funzione trova la posizione dei k minimi di un vettore

```
se.hat<-function(T,D,V,c,rho)
      {
      den<-sum((T>=c)*(V*D+(1-V)*rho))
      num<-sum(V*D+(1-V)*rho)
      se<-den/num
      se
      }
```

Qusta funzione stima la sensitività utilizzando lo stimatore proposto (1.5)

```
sp.hat<-function(T,D,V,c,rho)
      {
      den<-sum((T<c)*(V*(1-D)+(1-V)*(1-rho)))
      num<-sum(V*(1-D)+(1-V)*(1-rho))
      sp<-den/num
      sp
      }
```

Qusta funzione stima la specificità utilizzando lo stimatore proposto (1.6)

```
se.full<-function(T,D,c)
{

  tp<-sum(T[D==1]>c)
  tot<-sum(D==1)
  tp/tot
}
```

Questa funzione calcola la sensitività nel caso le osservazioni siano considerate tutte verificate (1.1).

```
sp.full<-function(T,D,c)
{
  tn<-sum(T[D==0]<c)
  tot<-sum(D==0)
  tn/tot
}
```

Questa funzione calcola la specificità nel caso le osservazioni siano considerate tutte verificate (1.2).

```
distanza<-function(x0,x1)
{
  xa<-x0[,1]
  xb<-x1[,1]
  ta<-x0[,2]
  tb<-x1[,2]
  a<-length(xa)
  b<-length(xb)
  dis<-matrix(0,a,b)
  for(i in 1:a)
  {
    x<-(xb-xa[i])^2
    t<-(tb-ta[i])^2
    dd<-sqrt(x+t)
    dis[i,]<-dd
  }
  dis
}
```

Questa funzione riceve due matrici A_{ax2} e B_{bx2}. Si considerino come coordinate di punti su un piano cartesiano le singole righe di entrambe la matrici. Per ogni punto appartenente alla matrice A la funzione restituisce la distanza euclidea da tutti i punti della matrice B.

```
covarianza<-function(X,Y)
{
  n<-length(X)
  H<-diag(n)-matrix(1,n,n)/n
  x<-cbind(X,Y)
  S<-(t(x)%*%H%*%x)/n
  S
}
```

Questa funzione calcola la matrice di varianza/covarianza per due vettori

```
distanzamaha<-function(T,X,l1,covar,n)
{
```

```
3    dis<-matrix(0,n,n)
4    covinv<-solve(covar)
5    for(i in 1:n)
6    {
7      for(j in 1:i)
8      {
9        a<-c(T[i]-T[j],X[i]-X[j])
10       dd<-sqrt(t(a)%*%covinv%*%a)
11       dis[j,i]<-dd
12     }
13
14   }
15   aa<-(l1+1)
16   distrid<-dis[1:l1,aa:n]
17   distrid
18 }
```

Siano X e T due vettori, questa funzione calcola la distanza di Mahalanobis tra le coppie $(X_i, Tt_i | i < l1)$ e le $(X_j, Tt_j | j > l1)$.

```
1  distanzamahn<-function(T,X,l1,n)
2  {
3    dis<-matrix(0,n,n)
4    for(i in 1:n)
5    {
6      for(j in 1:i)
7      {
8        dd<-abs(T[i]-T[j])+abs(X[i]-X[j])
9        dis[j,i]<-dd
10     }
11
12   }
13   aa<-(l1+1)
14   distrid<-dis[1:l1,aa:n]
15   distrid
16 }
```

Siano X e T due vettori, questa funzione calcola la distanza di Manhattan tra le coppie $(X_i, Tt_i | i < l1)$ e le $(X_j, Tt_j | j > l1)$.

```
distanzalag<-function(T,X,l1,n)
{
  dis<-matrix(0,n,n)
  for(i in 1:n)
  {
    for(j in 1:i)
    {
      dd<-max(abs(T[i]-T[j]),abs(X[i]-X[j]))
      dis[j,i]<-dd
    }

  }
  aa<-(l1+1)
  distrid<-dis[1:l1,aa:n]
  distrid
}
```

Siano X e T due vettori, questa funzione calcola la distanza di Lagrange tra le coppie $(X_i, Tt_i | i < l1)$ e le $(X_j, Tt_j | j > l1)$.

A.4 Grafici

```
library (sm)
par(mfrow=c(1,2))

 hist ( a[,8] , freq =FALSE , main = 'valori auc', sub =
    paste('n= ',n,' k= ', k,' alpha= ', alpha))
 points ( soluzauc[1], 0, pch="*", col=" red ", cex =3) #
    media Jackknife
 points ( full[5] , 0, pch="+", col=" blue ", cex =3) #
    media full
 points (soluzauc[3], 0, pch="o", col=" green ", cex =2)
    #media Monte Carlo
 legend ("topleft", legend =c('media JK', 'media full',
    'media MC') , col=c("red", " blue",
    'green'),pch=c('*','+','o'), pt.cex =c(2,2,1,5))
 h.hat <- h.select (a[,8] , method ="cv")
 sm.density (a[,8] , h=h.hat , add= TRUE )
```

```
11
12
13 hist ( a[,7] , freq =FALSE , main = 'varianza auc', sub
     = paste('n= ',n,' k= ', k,' alpha= ', alpha))
14 points ( soluzauc[2], 0, pch="*", col=" red ", cex =3)
     #varianza Jackknife
15 points ( soluzauc[4] , 0, pch="o", col=" blue ", cex =2)
     # varianza Monte Carlo
16 legend ("topright", legend =c('varianza JK', 'varianza
     MC') , col=c("red", " blue"),pch=c('*','o'), pt.cex
     =c(2,1,5))
17 h.hat <- h.select (a[,7] , method ="cv")
18 sm.density (a[,7] , h=h.hat , add= TRUE )
```

Bibliografia

[1] Adimari G., Chiogna M., 2012, *Nearest-neighbor estimation for ROC analysis under verification bias.*

[2] Zamengo B., 2013, *A Monte Carlo study of the properties of a nonparametric estimator for the area under the ROC curve.*