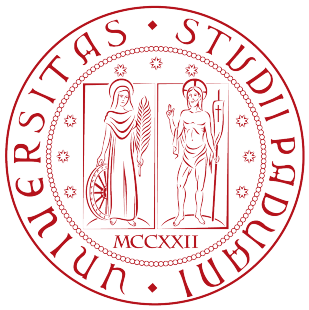
Università degli studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze



RELAZIONE FINALE

Studio statistico sul grado di accordo nella diagnosi del Papilloma Virus

Relatore Prof.ssa Laura Ventura
Dipartimento di Scienze Statistiche

Laureando: Maistri Pietro
Matricola N. 1198782

Anno Accademico 2021/2022

Indice

Introduzione

Il seguente lavoro ha lo scopo di osservare e analizzare il grado di accordo tra alcuni osservatori nelle valutazione e diagnosi del virus HPV, chiamato anche Papilloma Virus. Con il termine Papilloma Virus si fa riferimento a un ampia famiglia di virus, alcuni dei quali possono essere dannosi per la salute umana. La trasmissione avviene per via sessuale, la maggior parte delle volte comporta leggere infezioni passeggere, alcune infezioni possono però portare allo sviluppo di masse tumorali. Sopratutto nelle donne queste infezioni sono spesso correlate con i tumori del collo dell'utero.

A 15 osservatori è stato chiesto di valutare alcuni vetrini contenenti dei campioni rilevati tramite Pap Test. Il loro compito era quello di esprimere un giudizio su questo vetrino, indicando anche il livello di gravità del patogeno, se rilevato. Tale procedura è stata ripetuta dopo aver somministrato un training apposito sulla valutazione di vetrini istologici e risultati di Pap test. Tramite alcuni indici *agreement* si è andato poi a valutare il grado di accordo tra i 15 valutatori prima e dopo la somministrazione del training, per valutare il livello di competenza e formazione degli osservatori e l'efficacia o meno di questo training.

Dopo una introduzione al virus e ai dati, vengono presentati i due dataset che sono stati usati per le analisi e le relative variabili. Dopo un richiamo teorico sugli indici e le relative formule che verranno usati, si presenta un

analisi approfondita delle variabili. L'ultimo capitolo riguarda l'inferenza. Si cerca infine di rispondere alla domande se il training è stato efficace o meno nell'aumentare le conoscenze e la formazione degli osservatori nell'analisi dei risultati di Pap Test.

Capitolo 1

Il Papilloma Virus

Il termine Papilloma Virus o HPV (*Human Papilloma Virus*) si riferisce a una famiglia composta da oltre 100 varietà di virus diversi. Di questi virus, la maggior parte non rappresenta un pericolo per la salute delle persone che ne vengono a contatto. Tuttavia, una piccola quota di questi virus può comportare delle gravi conseguenze per l'organismo umano; questo virus, infatti, può evolversi verso una forma tumorale. Nello specifico il virus HPV16 e HPV18 sono responsabili del 70% dei tumori alla cervice uterina (Lega Italiana Lotta Ai Tumori, 2019).

1.1 L'evoluzione della ricerca e della prevenzione

Nell'ambito della ricerca e prevenzione delle malattie, l'uomo ha da sempre effettuato numerose ricerche per trovare test o metodi in grado di diagnosticare una malattia precocemente e se possibile evitare il suo sviluppo.
George Papanicolaou, ritenuto il padre della citopatologia (Tan e Tatsumura, 2015), nel 1943 fu il primo a proporre un test di screening per la preven-

zione del tumore alla cervice uterina, il test venne chiamato Pap test. Il test si divide in due fasi:

- prima viene prelevata una piccola quantità di cellule dal collo dell'utero tramite una spatola e un tampone cervicale. Le cellule vengono poi deposte su un vetrino per essere analizzate;

- in laboratorio un citopatologo analizza il vetrino in cerca del Papilloma Virus.

In Italia per refertare i vetrini da Pap Test viene usato il sistema Bethesda, ovvero il sistema per la compilazione standardizzata delle diagnosi di citopatologia cervicale (Solomon et al., 2002). Successivamente, si è arrivati ai più moderni HPV test. Le modalità di esecuzione sono le stesse del Pap test; tuttavia l'HPV test è risultato più sensibile e più efficace nel predire la possibilità di sviluppo di lesioni che potrebbero evolversi in tumori. E' però meno specifico, cioè identifica anche infezioni che potrebbero regredire spontaneamente senza portare complicanze (Lodi et al., 2021).
Attualmente in Italia, secondo le linee guida europee e della Commissione oncologica nazionale, nelle donne tra 25 e 65 anni, sarebbe opportuno effettuare un Pap test ogni 3 anni e un HPV test ogni 5 anni.
Negli ultimi anni, anche grazie alla diffusione di nuove tecnologie, per queste tipologie di studi si è passati sempre più a un approccio digitale tramite computer e scanner appositi.

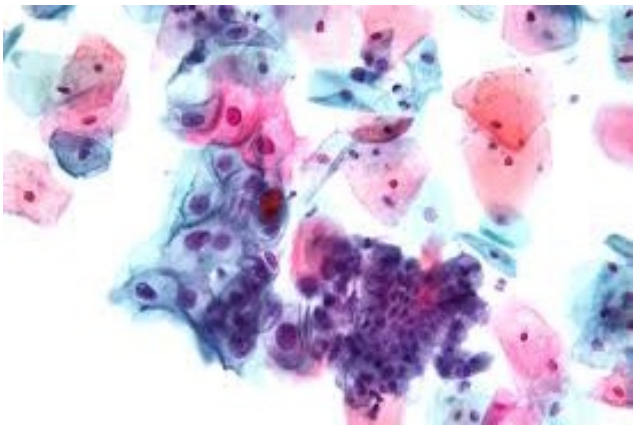


Figura 1.1: Esempio di un vetrino.

1.2 Il caso di studio

Il seguente lavoro ha lo scopo di valutare la conoscenza e il giudizio di alcuni citopatologi nello studio e screening del Papilloma Virus e l'accordo tra le diverse diagnosi. Nello specifico, tre esperti citopatologi hanno estratto 70 vetrini di uno screenging oncologico di un HPV test. I vetrini sono stati scannerizzati, resi anonimi e inseriti su un server online a cui si poteva accedere solo tramite password. Nella Figura 1.1 è riportato un esempio di un vetrino con presenza di virus HPV. E' stato quindi chiesto ad alcuni medici di valutare questi campioni per diagnosticare la presenza o meno del virus. In particolare a 15 esperti citopatologi, di cui 8 donne e 7 uomini, provenienti da 13 diversi paesi europei, è stato chiesto di valutare i risultati del test.
In un secondo momento, i medici selezionati hanno svolto un training volto a migliorare la loro conoscenza, esperienza riguardo il Papilloma Virus. Alla fine è stato riproposto il test sui vetrini di screening per vedere se il training ha migliorato la diagnosi.
Nella seguente relazione, dopo una presentazione dei dati e delle variabili,

varrà fatto un breve richiamo teorico su alcune misure di *agreement* utilizzate. Successivamente, si passerà a un' analisi esplorativa delle variabili d'interesse. Si proseguirà poi con una sezione dedicata all'inferenza statistica dove si proverà a quantificare l'efficacia o meno del training, basandosi sulle variabili a disposizione.

Capitolo 2

Presentazione dei dataset

Per il seguente lavoro sono stati utilizzati due *dataset*, vengono ora in seguito presnetati e descritti.

2.1 Primo dataset

Il primo *dataset* riguarda la valutazione dei vetrini da parte dei citopatologi. E' stato costruito tenendo conto sia della valutazione dei medici pre e post training, proprio per analizzare l'efficacia o meno di questo training. Ogni medico analizza un vetrino contenente o meno il virus, e indica un valore su una scala da 0 a 4, dove 0 indica una lieve presenza del virus e 4 una presenza molto grave del virus.
Sono state raccolte 70 osservazioni sui diversi medici e in tutto le variabili di cui si è tenuto conto sono 33, come riportato in Tabella 2.1 . Le variabili `obs` vanno da 1 a 15 e sono le valutazioni dei 15 citopatologi sui 70 diversi vetrini. Sono variabili discrete che hanno valori da 0 a 4, indicando la classificazione della patologia. In modo uguale le variabili `obsb`, che vanno sempre da 1 a 15 sono le valutazioni dei citopatologi dopo il training. Inoltre abbiamo:

Variabile	Definizione
`age`	Variabile che indica l'età del paziente osservato
`HPVstatus`	variabile dicotomica (Negative, Positive) che indica la presenza o assenza della malattia
`OrigianlCode`	variabile discrete che indica il vero valore del vetrino osservato (0-3)
`obs`	variabili discrete che indicano le 15 valutazioni degli osservatori sui vetrini (0-4) pre-training
`obsb`	variabili discrete che indicano le 15 valutazioni degli osservatori sui vetrini (0-4) post-training

Tabella 2.1: Variabili del primo dataset

Sarà importante analizzare in che modo cambia la valutazione degli osservatori prima e dopo il training, per valutare quanto questo sia stato efficace o meno.

2.2 Secondo dataset

Il secondo *dataset* invece riguarda alcune caratteristiche dei 15 osservatori ed è composto da 38 variabili per 39 osservazioni. Il numero di osservatori appare due volte perché si tiene conto sia della valutazione pre che post training sopratutto per calcolare alcune misure di associazione. Nel *dataset* viene riportata la correlazione di Spearman, di Kendall e il Kappa di Fleiss per valutare l'*agreement* tra la valutazione dell'osservatore e il vero valore

del vetrino (variabile `OriginalCode`).

Le variabili sono descritte nella Tabella 2.2 e nella Tabella 2.3.

Variabile	Definizione
`Observer`	variabile indicatrice dell'osservatore (0-15),
`Senior`	variabile dicotomica che indica se l'osservatore è *senior* o no(Un osservatore è ritenuto *senior* se ha più di 10 anni di esperienza),
`kend`	correlazione di Kendall
`time`	variabile indicatrice, 1 pre training, 2 post training,
`spearman`	correlazione di Spearman,
`practicepath`	variabile discreta indica la pratica (in anni) in patologia del osservatore,
`practicecyto`	variabile discreta indica la pratica (in anni) in cito-patologia,
`expThinPre`	variabile discreta che indica l'esperienza (in anni) con ThinPrep,
`expHPV`	variabile discreta che indica l'esperienza in Triage (in anni) riguardanti l'HPV,
`cytovol`	variabile che indica la percentuale di attività che l'osservatore dedica alla *cytopathology*,
`cytocases`	variabile continua che indica il numero di casi di *cytopathology* che l'osservatore ha trattato,
`work`	variabile discreta che indica l'ambiente di lavoro:`ah` ospedale accademico,`gh` ospedale generale, `oh` altro ospedali e `pl` laboratori privati,
`digpathexp`	variabile dicotomica, che indica l'esperienza in citopatologia digitale (1 si, 0 no)
`digpathexp1`	variabile dicotomica, che indica l'esperienza con vetrini istologici nel lavoro di routine di un osservatore (1 sì, 0 no)
`digpathexp2`	variabile dicotomica, che indica l'esperienza con vetrini istologici nel lavoro di ricerca (1 sì, 0 no)
`digpathexp3`	variabile dicotomica, che indica il l'esperienza con vetrini citologici nel lavoro di routine di un osservatore (1 sì, 0 no)
`digpathexp4`	variabile dicotomica, che indica il l'esperienza con vetrini citologici nel lavoro di ricerca di un osservatore (1 sì, 0 no)
`kappa`	Kappa di Cohen
`kappa1`	Kappa di Cohen pesato

Tabella 2.2: Variabili sugli osservatori

Oltre a queste variabili, sono stati registrati anche 17 item sugli osservatori riguardo le modalità con cui hanno analizzato le valutazioni dei vetrini e come hanno lavorato.

Variabile	Definizione
`item1`	Ha osservato i casi nei giorni lavorativi? (1 Sì, 0 no, 2 entrambi)
`item2`	Osservare i vetrini digitali richiedeva più tempo rispetto l'osservazione con microscopio? (Da 1 a 5, 1 molto meglio, 5 molto peggio)
`item3`	Guardare i vetrini è stato più stancante per gli occhi rispetto l'uso del microscopio? (Da 1 a 5, 1 molto meglio, 5 molto peggio)
`item4`	Guardare i vetrini è stato più stancante per la schiena rispetto l'uso del microscopio? (Da 1 a 5, 1 molto meglio, 5 molto peggio)
`item5`	Ci sono stati dei problemi con la nitidezza delle immagini? (1 Sì, 0 No)
`item6`	Quanto erano comuni i problemi di nitidezza dell'immagine? (Da 1 a 5, 1 ogni caso, 5 in nessun caso)
`item7`	Ci sono stati problemi con la qualità della colorazione dell'immagine? (1 Sì, 0 No)

Variabile	Definizione
item8	Quanto erano comuni i problemi con la qualità della colorazione dell'immagine? (Da 1 a 5, 1 ogni caso, 5 in nessun caso)
item9	Ci sono stati problemi con la qualità della scansione dell'immagine? (1 Sì, 0 No)
item10	Quanto erano comuni i problemi con la qualità della scansione dell'immagine? (Da 1 a 5, 1 ogni caso, 5 in nessun caso)
item11	Nei casi problematici, la tua valutazione era più grave delle diagnosi effettiva? (1 Sì, 0 No)
item12	Dopo questo lavoro, suggeriresti l'uso della citologia digitale per il lavoro di routine? (1 Sì, 0 No)
item13	Quanto tempo è passato (in settimane) tra il primo e il secondo round?
item14	Hai guardato solamente la R.O.I ? (Sì, 0 No)
item15	Hai guardato l'intera diapositiva scansionata (1 Sì, 0 No)
item11	Con quale dispositivo hai guardato le immagini? (Pc, Ipad)
item17	Hai usato un pc con uno schermo apposito per queste immagini? (1 Sì, 0 No)

Tabella 2.3: Item sugli osservatori

Di tutte queste variabili sarà interessante vedere quali incidono di più sulla valutazione da parte degli osservatori e il tipo di approccio che hanno utilizzato per lo studio di questi casi.

Nel prossimo capitolo si farà un breve richiamo sulle misure di *agreement* sopra elencate che sono state usate sia come variabili nel secondo *dataset* sia per valutare gli osservatori.

Capitolo 3

Richiamo sulle misure di *agreement*

Nel *dataset* sono presenti alcuni indici, come la correlazione di Spearman e il coefficiente di Kendall, il Kappa di Cohen e il Kappa di Cohen pesato, già calcolati sui dati; inoltre utilizzeremo anche il Kappa di Fleiss.
In questo capitolo vengono brevemente richiamati tali indici. L'interpretazione convenzionale di tali indici di *agreement*, accordo è il seguente: $<$0.00 indica assenza di *agreement*, 0-0.20 indicata una scarso livello di *agreement*, 0.21-0.40 una leggero *agreement*, 0.41-0.60 indica una moderato *agreement*, 0.61-0.80 indica un ottima *agreement* e 0.81-1 indica un perfetto *agreement* (Lands e Koch, 1977) .

3.1 Coefficiente di correlazione di Spearman

Il **coefficiente di correlazione di Spearman** è un indice di correlazione non parametrico usato per calcolare la correlazione tra due variabili X e Y. Per calcolarlo la procedura è la seguente:

1. Vengono assegnati i ranghi da 1 a n ai valori di X, tenendo anche conto di eventuali valori uguali. La stessa procedura viene ripetuta anche per la variabile Y.

2. Per ogni coppia (x_i, y_i) si calcola la differenza d_i tra i loro ranghi, per $i = 1, \ldots, n$.

3. Il coefficiente di correlazione di Speareman è:

$$r_s = \frac{1 - 6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}. \tag{3.1}$$

Il coefficiente r_s varia tra -1 e 1 indicando la direzione e la forza della correlazione (Doğan, 2018). Nel nostro caso la variabile `spearman` è costruita calcolando l'indice sulle valutazioni degli osservatori e la variabile `OriginalCode`, per tutti e 15 i valutatori prima e dopo il training.

3.2 Correlazione di Kendall

Come la correlazione di Spearman anche il **coefficiente di correlazione di Kendall** è un indice non parametrico, che misura la forza e la direzione della relazione fra due variabili quantitative o qualitative ordinali.
Rispetto all'indice di Spearman è più complicato da calcolare e richiede più passaggi, per questo motivo spesso viene preferito l'indice di Spearman. Tuttavia per campioni meno numerosi si consiglia di usare l'indice di Kendall. L'interpretazione di questo indice è la medesima di quello di Spearman; il risultato è compreso tra -1 (perfetta concordanza negativa) e +1 (perfetta concordanza positiva), un valore pari a 0 indica l'assenza di qualsiasi concordanza tra le variabili.
Esistono 3 indici di Kendall (*tau-a*, *tau-b*, *tau-c*): in generale il *tau-b* è quello

proposto di default della maggior parte dei software.

La formula di calcolo è la seguente :

$$\tau = \frac{n_{\mathrm{c}} - n_{\mathrm{c}}}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \tag{3.2}$$

dove:

- n_c = numero di casi concordanti,
- n_d = numero di casi non concordi,
- $n_0 = \frac{(n-1)}{2}$,
- $n_1 = \sum_i \frac{t_{\mathrm{i}}(t_{\mathrm{i}}-1)}{2}$, dove t_i è il numero di casi concordi nel gruppo i-esimo della variabile X,
- $n_2 = \sum_j \frac{u_{\mathrm{j}}(u_{\mathrm{j}}-1)}{2}$, dove u_j è il numero di casi concordi nel gruppo j-esimo della variabile Y.

Sia l'indice di Kendall che l'indice di Spearman sono indici non parametrici e rispetto alla correlazione di Pearson non richiedono assunzioni sulla distribuzione dei dati (Muliere, 1976).
Nel secondo dataset la variabile `kend` è stata costruita usando questo indice per ogni osservatore sulla sua valutazione (`obs1`, ..., `obs15`).

3.3 Kappa di Fleiss

La statistica **Kappa di Fleiss**, proposta da Fleiss nel 1971, è un indice tra i più utilizzati per misurare l'accordo tra vari esaminatori con risposte categoriche. Tale indice è una generalizzazione dell'indice Kappa di Cohen del 1960.

La statistica è la seguente

$$K = \frac{\overline{P} - \overline{P_{\mathrm{e}}}}{1 - \overline{P_{\mathrm{e}}}}, \tag{3.3}$$

dove:

- N è il numero di soggetti,
- n è il numero di esaminatori,
- M è il numero di categoria di classificazione,
- x_{ij} sono il numero di esaminatori che hanno assegnato l'i-esimo soggetto $(i = 1, \ldots, N)$ alla j-esima categoria $(j = 1, \ldots, M)$,
- $P_{ij} = \frac{x_{ij}(x_{ij}-1)}{n(n-1)}$,
- $P_i = \sum_j p_{ij}$,
- $\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i$,
- $p_{ij} = \frac{1}{Nn} \sum_i x_{ij}$,
- $\bar{P}_e = \sum_{j=i}^{n} p_{ji}^2$

Questa misura di concordanza varia da 0 (assenza di concordanza) a 1 (concordanza perfetta) (Mitani et al., 2017).

3.4 Kappa di Cohen

Il **kappa di Cohen** è una misura dell'accordo tra le risposte qualitative o categoriali di due persone (*inter-observer variation*) oppure della medesima persona in differenti momenti (*intra-observer variation*), valutando gli stessi

soggetti. Questa statistica è stata presentata nel 1960 da Jacob Cohen (Nelson e Edwards, 2008).
La formula generale è la seguente:

$$k = \frac{p_0 - p_c}{1 - p_c}, \tag{3.4}$$

dove:

- p_0 è la proporzione dell'accordo osservato,
- p_c è la proporzione dell'accordo casuale.

Questa statistica varia fa -1 a 1, dove 1 indica un completo accordo, -1 un completo disaccordo e 0 nessun tipo di accordo tra le variabili. Di questa indice è disponibile anche una versione "pesata" ovvero che attribuisce pesi (gravità) diversi agli errori di classificazione. Secondo questa logica, il disaccordo nell'attribuzione di un' unità a due categorie differenti è da ritenere più grave quanto più le categorie sono distanti tra loro nella scala ordinale. Viene riportata ora la formula di calcolo:

$$k = \frac{p_{w0} - p_{wc}}{1 - p_{wc}}, \tag{3.5}$$

dove:

- p_{w0} è la proporzione pesata dell'accordo osservato,
- p_{wc} è la proporzione pesata dell'accordo casuale.

Di per sé il Kappa di Cohen non è quindi nient'altro che un caso particolare del Kappa pesato in cui i pesi sono tutti uguali a 0 (Mitani et al., 2017). Nel nostro dataset abbiamo in `kappa` il Kappa di Cohen e in `Kappa1` la versione pesata.
Si analizzano nel seguito tutte le variabili presentate in entrambi i due dataset oggetti di studio.

Capitolo 4

Analisi esplorative

In questo capitolo si presenta una prima analisi di alcune delle variabili a disposizione, calcolando alcune statistiche d'interesse. In alcuni casi si farà uso di alcune tecniche di simulazione via bootstrap per valutare la deviazione standard di alcune statistiche. Infine si proverà a quantificare le differenze nelle valutazioni, rispetto a quello che è il vero valore dei vetrini.

4.1 Analisi primo dataset

Da una prima analisi sul primo *dataset* possiamo notare come nel nostro campione abbiamo 46 casi di positività al virus e 24 casi negativi. L'età dei pazienti va dai 20 ai 79 anni, la media è 40.34 anni e la deviazione standard pari a 14.59.
Nella Figura 4.1 viene riporta la distribuzione dei veri valori dei vetrini codificati nella variabile `OriginalCode`.

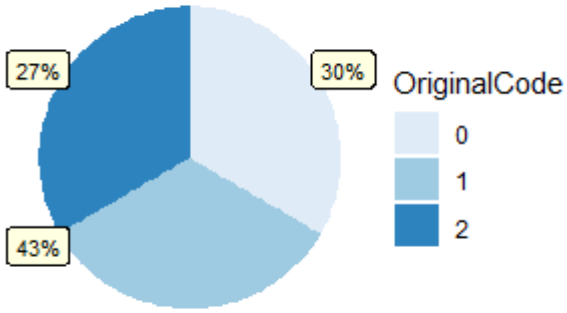


Figura 4.1: Grafico a torta variabile `Original Code`

Si ricorda che il valore 0 indica una lieve gravità del virus, mentre 2 indica un elevata presenza del virus.
Nel boxplot in Figura 4.2 possiamo osservare come cambiano le valutazioni degli osservatori dopo il training.

Notiamo subito come la mediana resti costante nei gruppi indipendente dal training. Salvo alcuni osservatori non sembra esserci un evidente effetto del training sulle valutazioni, i boxplot a coppie risultano simili nella maggior parte dei casi.
Andiamo ora a valutare tramite il Kappa di Fleiss e la correlazione di Kendall l'`agreemnt` tra le diverse modalità delle variabili, per valutare la differenza trai due indici calcolati useremo il Test di Cocron per maggiori approfondimenti su questo testi si rimanda a Cronbach (1951). Valutando il Kappa di Fleiss, troviamo un il valore 0.287 prima e 0.306 dopo il training, questi valori non sono distanti da loro L'indice di Fleiss calcolato sugli osservatori

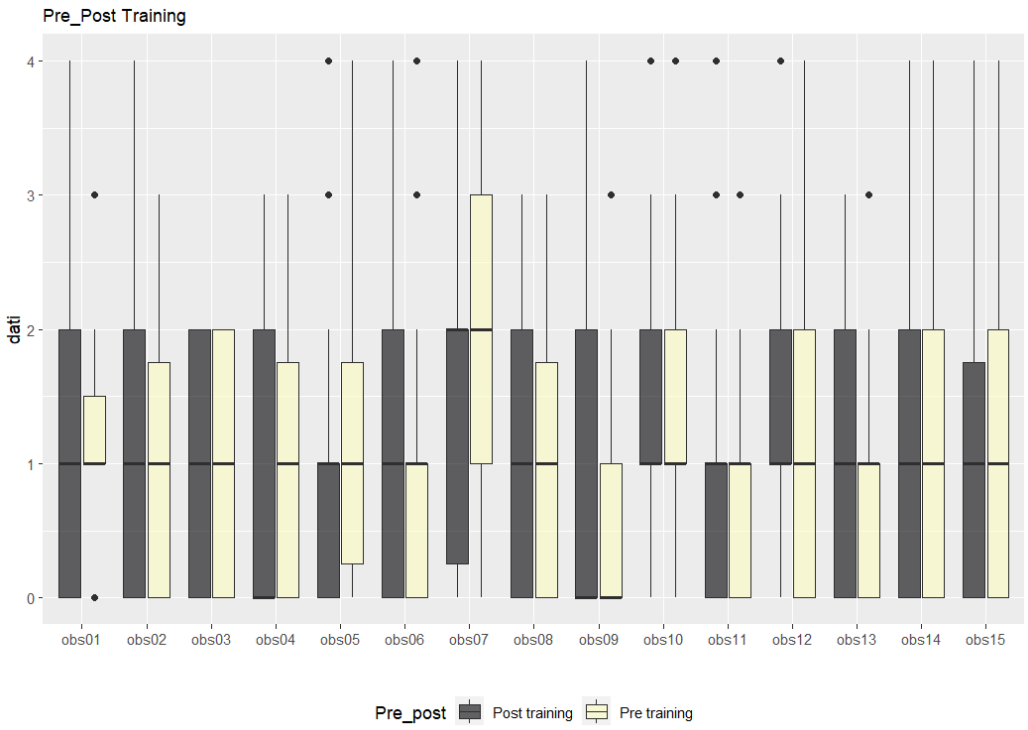


Figura 4.2: Boxplot pre e post training

pre e post training suggerisce una leggera concordanza (0.287 e 0.306).
Se andiamo invece a dividere gli osservatori in base all'esperienza, ovvero gli osservatori con più di 10 anni di esperienza (chiamati Senior) e quelli con meno di 10 anni di esperienza (chiamati Junior), il valore della statistica Kappa di Fleiss tra gli osservatori Senior è 0.247 mentre per i Junior 0.22, la differenza non è così elevata. Se dividiamo i casi pre e post training, per i Senior la statistica vale 0.298 pre e 0.313 post, per i Junior invece vale 0.247 pre e 0.273 post. Calcoliamo in modo analogo l'indice di Kendall: per i Senior vale 0.41 (0.535 pre e 0.528 post), mentre per i Junior vale 0.385 (0.442 pre e 0.52 post). Questi valori ci suggeriscono una concordanza positiva tra i valutatori, la concordanza è maggiore negli osservatori Senior, questo è probabilmente dovuto alla loro esperienza; inoltre, si nota come post training

le statistiche di *agreement* aumentino, tranne per la correlazione di Kendall negli osservatori Senior dove si nota un leggero calo (da 0.535 a 0.528).
Introduciamo ora l'analisi tenendo conto della variabile `HPVstatus`, che ci fornisce l'informazione se il paziente era positivo o meno al virus, calcoliamo gli indici di concordanza prima del training. D'ora in poi quando verranno calcolate le stime delle deviazioni standard dei vari indici si ricorda che tali stime sono ottenute via simulazione con un bootstrap non parametrico.
Prima del training troviamo che il Kappa nei positivi è 0.292 mentre per i negativi 0.261, i due indici hanno deviazione standard, rispettivamente, 0.042 e 0.065. Il test di Cocron accetta l'ipotesi di uguaglianza dei due valori per ogni livello di α usuale ($_{value} = 0.71$).

Per l'indice di Kendall invece le stime sui negativi è 0.492 e 0.524 per i positivi, le deviazioni standard ottenete via bootstrap sono rispettivamente 0.077 e 0.053. Anche in questo caso accettiamo l'ipotesi di uguaglianza dei due indici ($_{value} = 0.571$.)
Per quanto riguarda il post training, il Kappa di Fleiss vale 0.269 per i negativi (con deviazione standard 0.065) e per i positivi è 0.284 (deviazione standard 0.037), i coefficienti di Kendall sono rispettivamente 0.519 calcolati sui negativi (deviazione standard 0.095) e 0.545 per i positivi (deviazione standard pari a 0.049). In entrambi i casi il test di Cocron accetta l'ipotesi di eguaglianza dei due indici.
In questo caso i risultati sono molto simili, non sembra molto influente nella valutazione delle misure di *agreement* il fatto che un paziente sia positivo o negativo, tale variabile non influenza la valutazione.
Se invece andiamo a valutare le varie stime facendo riferimento al training, per tutti gli indici calcolati accettiamo l'ipotesi di uguaglianza tra l'indice pre e l'indice post, i risultati sono riportati nella Tabella 4.1.

	Pre	Post	$value$
$Kappa_{negativi}$	0.26	0.27	0.94
$Kendall_{negativi}$	0.49	0.51	0.70
$Kapppa_{positivi}$ 3	0.29	0.28	0.92
$Kendall_{positivi}$ 4	0.52	0.54	0.65

Tabella 4.1: Confronto indici Kappa di Fleiss e Kendall tra positivi e negativi prima e dopo il training

Andiamo a studiare ora come variano gli indici in funzione della variabile `OrigianlCode` che ricordiamo indica il vero valore del vetrino, sempre facendo distinzione tra il pre e il post training. Se guardiamo i casi prima del training, in cui `OrigianlCode` vale 0, troviamo un valore di Kappa pari a 0.317 (deviazione standard 0.053), quando il vero valore del vetrino è 1 il Kappa vale 0.189 (deviazione standard 0.043), quando il valore è 2, Kappa vale 0.26 (deviazione standard 0.056). Le stime della correlazione di Kendall sono invece 0.616 per `OrigianlCode` pari a 0 (deviazione standard 0.089), 0.42 (deviazione standard 0.056) quando è uguale a 1, 0.442 (deviazione standard 0.067), quando è uguale a 2. Per il post training invece abbiamo Kappa 0.284 (deviazione standard 0.07) quando `OrigianlCode` vale 0, Kappa 0.195 (deviazione standard 0.034) quando vale 1 e Kappa 0.155 (deviazione standard 0.038) quando vale 2. Le correlazioni di Kendall valgono, rispettivamente, 0.459 per `Originalcode` pari a 0, 0.516 per `Originalcode` pari a 1 e 0.356 per `Originalcode` pari a 2, le relative deviazioni standard sono 0.084, 0.058 e 0.06. Notiamo che nel pre training vi è un maggiore accordo tra gli osservatori quando il valore del vetrino è 0, forse per una maggior facilità a valutare il campione, questo effetto rimane anche dopo il training ma in maniera meno evidente. Infine notiamo che quando abbiamo il valore

0 della variabile `OrigianelCode` l'accordo è maggiore prima del training, rispettivamente Kappa pari a 0.317 e Kendall 0.616, dopo il training abbiamo Kappa 0.284 e Kendall 0.459, per `OriginalCode` pari a 1 l'effetto è minore, i valori di Kappa restano vicini, passano da 0.189 a 0.195 ma aumenta l'indice di Kendall, da 0.42 a 0516, mentre quando `OriginalCode` vale 2, il Kappa passa da 0.26 a 0.155 e Kendall diminuisce da 0.442 a 0.356, pare quindi non esserci un effetto significativo sul grado di accordo tra gli osservatori del training. Aiutandoci con il test di Cocron andiamo a valutare le differenze tra gli indici prima e dopo il training, i risultati sono riportati tutti nella Tabella 4.2.

	Pre	Post	$_{value}$
$Kappa_{OriginalCode=0}$	0.32	0.28	0.75
$Kappa_{OriginalCode=1}$	0.19	0.20	0.95
$Kappa_{OriginalCode=2}$	0.26	0.15	0.00
$Kendall_{OriginalCode=0}$	0.62	0.46	0.02
$Kendall_{OriginalCode=1}$	0.42	0.52	0.15
$Kendall_{OriginalCode=2}$	0.44	0.36	0.37

Tabella 4.2: Confronto indici Kappa di Fleiss e Kendall con la variabile `OrigianlCode` prima e dopo il training

Da questi confronti si nota che l'ipotesi di uguaglianza degli indici è accetta per la maggior parte dei casi, gli unici due casi dove pare esserci un deciso effetto del training è nell'indice di Kendall quando `OriginalCode` ha valore 0 e nel Kappa di Fleiss quando `OrigianalCode` vale 2, in tutti gli altri casi non si nota un effetto training poiché gli indici non sono significativamente diversi tra loro. Andiamo ora a ad analizzare le singole valutazione degli osservatori e vedere come in media si avvicinano ai veri valori.

4.1.1 Analisi delle differenze tra le valutazioni e la variabile `OriginalCode`

Andiamo ora a calcolare la differenza tra le valutazioni degli osservatori e la variabile `OriginalCode`, per valutare quante buone sono le valutazioni dei vetrini assegnate dagli osservatori con quello che rappresenta il vero livello di "gravità". Si ricorda che più la valutazione è elevata più è grave la presenza di virus sul campione. Osserviamo prima i risultati pre training, riportati nella Tabella 4.3.

	Media	Deviazione standard
obs 1	0.05	0.92
obs 2	0.00	0.90
obs 3	0.07	0.60
obs 4	-0.14	0.91
obs 5	0.23	1.23
obs 6	-0.03	1.02
obs 7	0.86	1.39
obs 8	-0.01	0.99
obs 9	-0.26	1.08
obs 10	0.36	1.08
obs 11	-0.13	1.03
obs 12	0.09	1.22
obs 13	-0.07	0.91
obs 14	0.04	1.20
obs 15	0.07	1.15

Tabella 4.3: Tabelle differenze pre training

La situazione ideale è quella in cui tutti gli osservatori hanno una differenza media pari a 0, ovvero in media le valutazioni coincidono con i veri valori. Nel nostro caso vediamo come gli osservatori che danno i migliori giudizi sono gli osservatori 1, 2, 6, 8, 14, le loro medie non si discostano di ± 0.05 dal valore ideale 0. Gli osservatori con le peggiori valutazioni sono invece sono gli osservatori 7, 9 e 10, la media delle loro osservazioni è più distante da 0. In generale la differenza media pre training è 0.074 e la deviazione standard media è pari a 1.041. Vediamo come cambia la situazione dopo il training nella Tabella 4.4.

	Media	Deviazione standard
osb 1	-0.04	0.95
osb 2	0.10	0.83
osb 3	0.17	0.66
osb 4	-0.24	0.97
osb 5	-0.13	0.93
osb 6	0.07	0.95
osb 7	0.74	1.30
osb 8	-0.04	0.86
osb 9	-0.06	1.14
osb 10	0.31	0.91
osb 11	-0.06	1.14
osb 12	0.49	1.15
osb 13	0.03	0.78
osb 14	0.06	1.19
osb 15	-0.13	0.90

Tabella 4.4: Tabelle differenze post training

Ora i risultati sono diversi rispetto a quelli precedenti. Gli osservatori migliori sono il numero 1, 8, 13. Se guardiamo quelli che prima erano risultati essere i peggiori osservatori, notiamo un lieve miglioramento in media per l'osservatore 7 (da 0.86 a 0.74) e l'osservatore 10 (da 0.36 a 0.31) e un netto miglioramento per l'osservatore 9 (da -0.26 a -0.06). Tuttavia la differenza media peggiora, da 0.074 a dopo il training 0.085, mentre la deviazione standard diminuisce di poco, 0.978. In sintesi, possiamo dire che le valutazioni degli osservatori portano dei buoni risultati già prima del training, il loro livello di competenze e formazioni su questo argomento era già elevato, il training pare non evidenziare miglioramenti netti nelle valutazioni.

4.2 Analisi secondo dataset

Analizziamo ora le variabili del secondo dataset, dove possiamo trovare alcune importanti informazioni sugli osservatori e sulla loro esperienza. Dei 15 osservatori in totale, 7 di loro sono Senior mentre 8 sono Junior. Ricordiamo che un osservatore per potersi definire Senior deve avere più di 10 anni di esperienza.
Nel boxplot in Figura 4.3 possiamo notare le distribuzione delle variabili relative agli anni di pratica degli osservatori in *pathology* e in *cytopathology*. Le medie per le due variabili è uguale 24.20 anni, ma la pratica in *cytopathology* ha una maggior deviazione standard (8.58) rispetto la pratica in *pathology* (8.09). Nel boxplot in Figura 4.4 osserviamo le variabili relative all'esperienza con *Thin Prep* (tipologia di analisi dei vetrini per malattie come il Papilloma Virus), dati relativi ai virus HPV, con *triage* relativi all'HPV e in attività di screening primario.

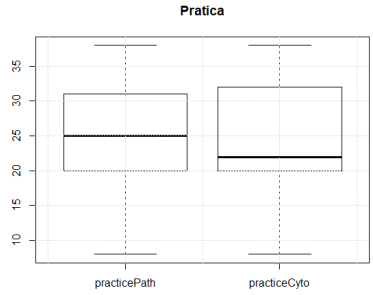


Figura 4.3: Boxplot pratica

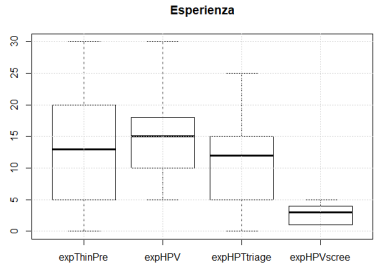


Figura 4.4: Boxplot esperienza

Notiamo che tuttavia gli osservatori sembrano avere esperienza elevata per quanto riguarda le prime tre variabili, mentre questa è minore quando si parla di attività di screening primario per l'HPV. Nella Tabella 4.5 sono riportate le correlazioni tra queste variabili.

	expThinPre	expHPV	expHPVtriage	expHPVscree
expThinPre	1	0.6445	0.5005	0.2754
expHPV	0.6445	1	0.7017	0.0567
expHPVtriage	0.5005	0.7017	1	0.1014
expHPVscree	0.2754	0.0567	0.1014	1

Tabella 4.5: Correlazioni esperienza

Se osserviamo le correlazioni notiamo che le prime tre le variabili hanno una buona correlazione tra di loro, mentre per quanto riguarda `expHPVscree` i valori sono minori, 0.275, 0.0567 e 0.1; questo può essere dovuto al fatto che solitamente ci sono altre figure negli ambienti ospedaliero-sanitario volte a occuparsi specificamente delle attività di screening delle malattie. Se osserviamo la variabile `kend` possiamo notare i diversi valori per la correlazione di Kendall pre e post training, calcolata tra le valutazioni degli osservatori

e il vero valore del vetrino sulle variabili del primo *dataset* i risultati sono riportati nalla Tabelle 4.6.

	Kendall Pre	Kendall Post
osb 1	0.86	0.89
osb 2	0.87	0.89
osb 3	0.89	0.83
osb 4	0.88	0.88
osb 5	0.88	0.87
osb 6	0.89	0.90
osb 7	0.93	0.92
osb 8	0.88	0.89
osb 9	0.88	0.88
osb 10	0.88	0.87
osb 11	0.88	0.88
osb 12	0.89	0.91
osb 13	0.88	0.88
osb 14	0.89	0.90
osb 15	0.80	0.88

Tabella 4.6: Tabelle `kend`

Osservando la prima colonna della Tabella 4.6 notiamo subito delle correlazioni elevate per tutti gli osservatori, il risultato già prima del training pare soddisfacente, la correlazione media pre training è 0.878, mentre la deviazione standard è 0.027. Dopo il training le correlazioni migliorano leggermente: la media ora è 0.885 e la deviazione standard 0.022. Pare quindi che gli osservatori avessero già delle ottime conoscenze e abilità nel valutare i vetrini

già prima di fare essere sottoposti a questo training, in entrambi i casi si può dire che c'è un elevata concordanza positiva.
Osserviamo ora la variabile `spearman`, che indica proprio la correlazione di Spearman calcolata sulle variabili del primo *dataset*, nella Tabella 4.7.

	Spearman Pre	Spearman Post
osb 1	0.29	0.52
osb 2	0.32	0.59
osb 3	0.72	0.67
osb 4	0.43	0.45
osb 5	0.18	0.49
osb 6	0.37	0.51
osb 7	0.21	0.34
osb 8	0.19	0.47
osb 9	0.12	0.37
osb 10	0.34	0.53
osb 11	0.23	0.26
osb 12	0.20	0.38
osb 13	0.34	0.56
osb 14	0.30	0.37
osb 15	0.35	0.49

Tabella 4.7: Tabella `spearman`

In questa variabile è più forte l'effetto del training, la maggior parte degli osservatori dopo il training riporta una correlazione superiore rispetto a prima, difatti la correlazione media passa da 0.306 a 0.4667, con le deviazioni standard rispettivamente pari a 0.143 e 0.107. I risultati non cambiano se

osserviamo la statistica Kappa di Cohen, se osserviamo la variabile `kappa` i risultati sono analoghi a quelli precedenti, come riporta la Tabella 4.8.

	Kappa Pre	Kappa Post
osb 1	0.15	0.23
osb 2	0.17	0.42
osb 3	0.72	0.58
osb 4	0.16	0.28
osb 5	0.04	0.22
osb 8	0.06	0.36
osb 9	0.04	0.17
osb 6	0.16	0.18
osb 7	0.15	0.27
osb 10	0.22	0.26
osb 11	0.08	0.19
osb 12	0.23	0.23
osb 13	0.22	0.36
osb 14	0.24	0.15
osb 15	0.21	0.31

Tabella 4.8: Tabella `kappa`

In questo caso però i valori delle correlazioni sono più modesti, si nota un aumento post training, in media le correlazioni passano da 0.19 a 0,28, la deviazione standard invece diminuisce da 0.162 a 0.113.
Se nello stesso modo analizziamo la versione pesata del Kappa di Cohen, contenuta nella variabile `kappa1`, osservando la Tabella 4.9, notiamo una maggior effetto del training, la media pre training è di 0.27 (deviazione stan-

dard 0.15) mentre dopo il training la media diventa 0.41 (deviazione standard 0.12), quindi si passa da un accordo debole a un accordo già più marcato dopo il training.

	Kappa1 Pre	Kappa1 Post
osb 1	0.27	0.46
osb 2	0.32	0.53
osb 3	0.71	0.65
osb 4	0.42	0.40
osb 5	0.15	0.43
osb 6	0.33	0.46
osb 7	0.13	0.23
osb 8	0.19	0.45
osb 9	0.11	0.34
osb 10	0.27	0.41
osb 11	0.19	0.18
osb 12	0.15	0.30
osb 13	0.33	0.53
osb 14	0.23	0.31
osb 15	0.30	0.45

Tabella 4.9: Tabella `kappa1`

Rispetto quindi alla versione classica del Kappa di Cohen, le versione pesata segnala una maggior efficacia del training, per fare un confronto la media post training della misura classica è 0.28 mentre nella versione pesata è 0.41. Nelle figure 4.5, 4.6, 4.7 e 4.8 sono riportati i boxplot delle 4 distribuzioni.

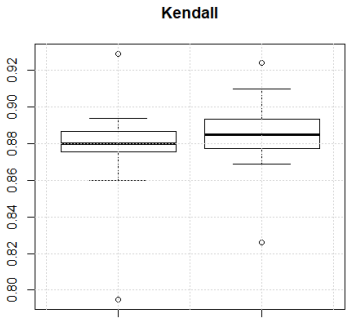


Figura 4.5: Boxplot Kendall

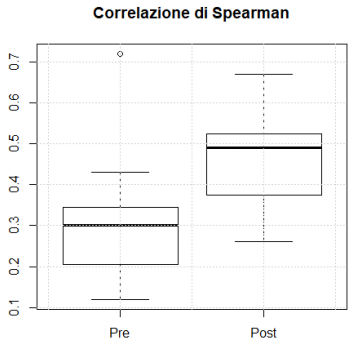


Figura 4.6: Boxplot Spearman

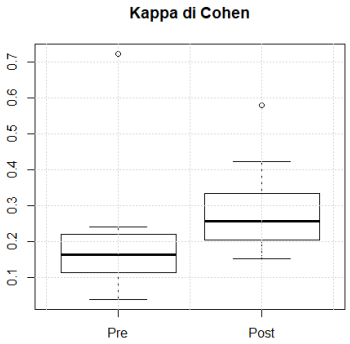


Figura 4.7: Boxplot Kappa

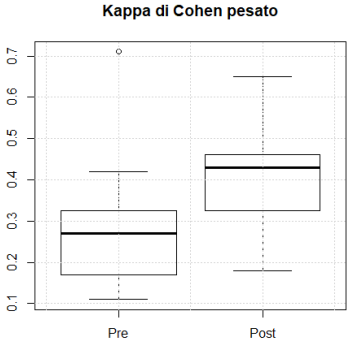


Figura 4.8: Boxplot Kappa pesato

Per ognuna delle 4 variabili, andremo ora a calcolare con un test se è ragionevole assumere che le distribuzioni pre e post training siano le medesime. A questo scopo useremo il Test di Mann-Whitney (Harris e Hardin, 2013). Per quanto riguarda la variabili `kend`, il test accetta l'ipotesi di uguaglianza delle distribuzioni, ($p_{value} = 0.37$), mentre per la correlazione di Spearman queste ipotesi viene rifiutata, ($p_{value} = 0.0008$). Vi è quindi una significativa differenze delle distribuzioni prima e dopo. Infine anche per le variabili `kappa` e `kappa1` possiamo affermare che c'è una differenza significative nelle loro distribuzioni (p_{value} pari a 0.005 e 0.004). Abbiamo quindi notato una

significativa risposta positiva degli osservatori al' training, tranne per l'indice di Kendall le variabili registrano tutte un maggior `agreement`.
Andiamo ora a studiare le variabile `work`, che indica l'ambiente lavorativo principale dell'osservatore.

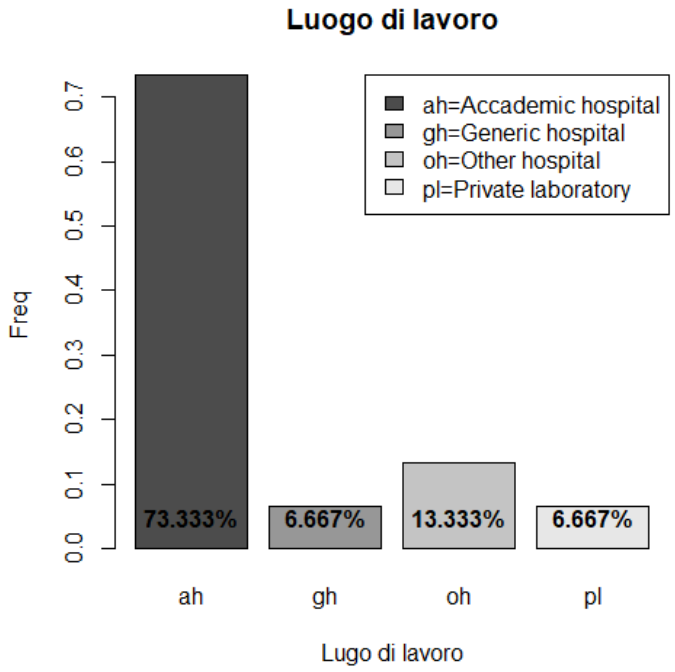


Figura 4.9: Istogramma `work`

Nella Figura 4.9 notiamo come la maggior parte dei nostri osservatori opera in ambienti ospedalieri accademici.
Per quanto riguarda le variabili sull'esperienza possiamo affermare che l'87% aveva già esperienza con la *digital cytopatologhy*, tuttavia solo il 27% aveva già esperienza con i vetrini istologici nel loro lavoro di ricerca, ma l'80% di loro aveva comunque utilizzato i vetrini istologici nel lavoro di ricerca. Risultati molto simili anche per quanto riguarda l'uso di vetrini citologici, solamente il 13% di loro aveva già utilizzato queste modalità nel loro lavoro

di routine, ma per quanto riguarda la ricerca queste tecniche erano già state utilizzate da circa l' 80% degli osservatori. Questo può essere dovuto al fatto che queste tecniche di analisi dei vetrini istologici e citologici sono molto più diffuse nell'ambito accademico e nel campo della ricerca, rispetto ad altri settori.

Studiamo ora le variabili relative alla quantità di lavoro che ogni osservatore dedica alla *cytopatholghy*, rispettivamente la variabile `cytovol` che indica la percentuale di lavoro dedicata a *cytopatholghy* e `cytocases` che è il numero totale di casi trattati. Nella Tabella 4.10 sono riportate alcune misure di sintesi della prima variabile.

Min.	Mediana	Media	Max	Dev. std
0	30	33.1	70	23.74

Tabella 4.10: Statistiche `cytovol`

Possiamo quindi dire che mediamente circa un terzo del lavoro del lavoro degli osservatori è dedicato a casi di studio legati citopatologia. Nella Figura 4.10 è rappresentato il boxplot di questa variabile.

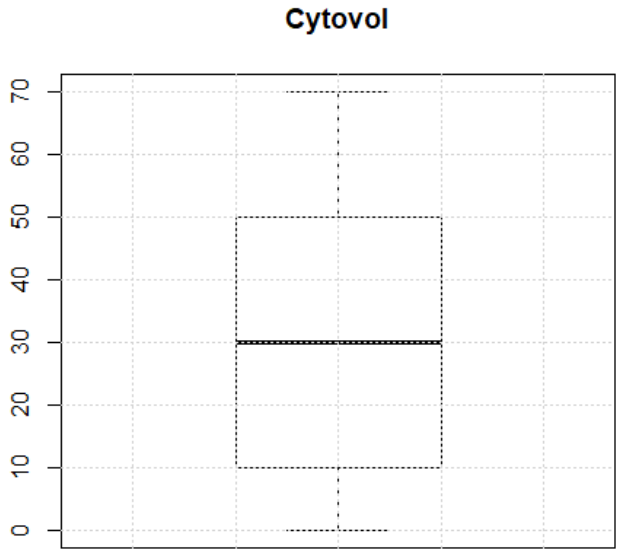


Figura 4.10: Boxplot `cytovol`

Per la variabile `cytocases` che indica il numero di casi totali di studio per ogni osservatore di citopatologia, il meno esperto ha solamente un totale di 300 casi di studio e il più esperto ha 15000 casi di studio, con una media poco inferiore ai 6000 casi di studio per ogni osservatore. Ulteriori misure sono riportate nella Tabella 4.11. Nella Figura 4.11 è riportato il boxplot della variabile.

Min.	Mediana	Media	Max	Dev. std
300	4500	5967	15000	4957.1

Tabella 4.11: Statistiche `cytocases`

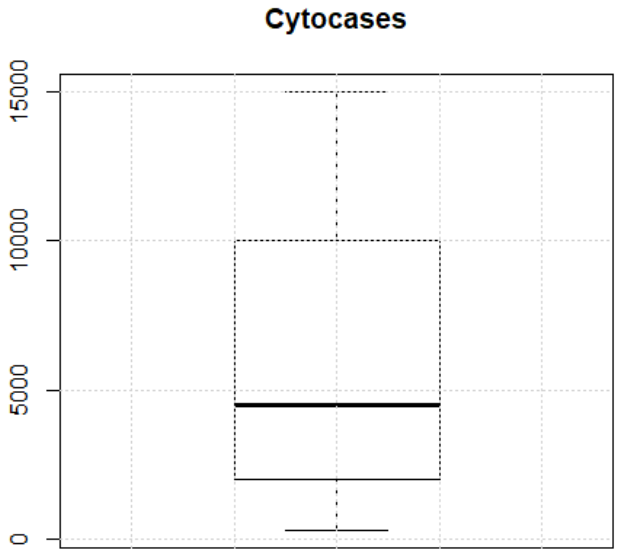


Figura 4.11: Boxplot `cytocases`

4.2.1 Analisi degli `item` raccolti sugli osservatori

Andando a studiare anche gli `item` raccolti sugli osservatori e le modalità con cui hanno lavorato, si può notare come il 60% degli osservatori ha lavorato ai vetrini solamente in settimana, i restanti hanno lavorato nel weekend o in entrambi i periodi. La maggior parte degli osservatori si è trovata molto bene nello studio dei vetrini in formato digitale piuttosto che con l'uso del microscopio. Tuttavia circa un terzo degli osservatori ha affermato che osservare i vetrini fosse più stancante per gli occhi e sopratutto per la schiena, rispetto a tecniche più classiche. Tuttavia il 94% degli osservatori ha dichiarato di aver avuto problemi con la nitidezza delle immagini visualizzate, questi problemi erano molto comuni nelle diapositive dei vetrini visualizzati. Per quanto ri-

guarda i problemi relativi alla colorazione delle immagini, 8 osservatori non dichiarano di averne avuti mentre 7 osservatori hanno avuto problemi con con la colorazione. Inoltre ben 12 osservatori su 15 hanno notato problemi con la qualità della scansione dell'immagine. Il 60% degli osservatori inoltre nei casi problematici ha dato una diagnosi più grave di quella effettiva. Nella domanda "suggeriresti l'uso della citologia digitale anche per il lavoro di routine", c'è stata una netta divisione tra gli osservatori, 7 di loro sono a favore mentre 8 di loro sono a sfavore. Inoltre sappiamo che tra la prima valutazione e la seconda sono passate dalle 2 alle 8 settimane, con una media di 5 settimane. Nella valutazione dei vetrini tutti gli osservatori hanno lavorato con un computer salvo 1 osservatore che ha visualizzato le immagini su un Ipad, tuttavia solo 3 osservatori disponevano di uno PC con schermo apposito per la riproduzione e analisi di queste immagini.
In conclusione, queste variabili ci fornisco delle importanti informazioni. Molti osservatori hanno riscontrato delle difficoltà con le immagini da valutare, per la nitidezza, la colorazione e la qualità. Inoltre pochi erano dotati degli strumenti più predisposti per queste tipologie di immagini, le valutazione potrebbero essere state influenzate da queste problematiche, non avendo un qualità ottimale delle vetrini, per gli osservatori può essere stato più difficile riscontrare o meno la presenza e la gravità del virus, questo può essere un problematica non da poco e portare a una forte distorsione dei risultati.
Questi dettagli vanno tenuti bene in considerazione nel caso si volesse ripetere lo studio in un futuro. Per evitare una possibile fonte di distorsione nelle valutazioni gli osservatori devono poter visualizzare e analizzare le immagini dei vetrini nel miglior modo possibile, con i migliori strumenti adatti a questo tipo di lavoro.

4.2.2 Analisi tre le variabili degli indici di *agreement* e le covariate

Sia andrà ora a calcolare la correlazione (usando la correlazione di Spearman) tra le variabili `kend`, `kappa`,`kappa1`, `spearman` e alcune covariate. I risultati sono riportati nelle Tabelle 4.12 e 4.13.

	kend	spearman	kappa	kappa1
practicePath	0.51	-0.09	0.13	-0.14
practiceCyto	0.45	0.04	0.19	0.00
expThinPre	0.19	0.31	-0.11	0.30
expHPV	0.24	0.13	0.20	0.22
expHPVtriage	0.45	0.02	0.19	0.08
expHPVscree	0.48	-0.13	-0.15	-0.23
cytovol	-0.16	0.25	0.42	0.32
cytocases	0.12	0.48	0.39	0.33

Tabella 4.12: Correlazioni Pre Training

Di queste covariate nessuna spicca per una correlazione elevata per tutti i nostri indici. La correlazione più rilevante riguarda la variabile `kend`, che ha una buona correlazione con le variabili relative alla pratica in citopatologia e patologia. Inoltre ha anche una correlazione con le variabili sull'esperienza nei *triage* medici e nelle attività di screening.

	kend	spearman	kappa	kappa1
practicePath	0.53	-0.15	-0.34	-0.19
practiceCyto	0.26	0.07	-0.15	-0.03
expThinPre	0.25	0.16	0.12	0.23
expHPV	0.32	0.39	0.23	0.47
expHPVtriage	0.23	0.09	-0.34	0.12
expHPVscree	0.20	-0.45	-0.02	-0.39
cytovol	-0.25	0.35	0.18	0.32
cytocases	0.00	0.47	0.38	0.36

Tabella 4.13: Correlazioni Post Training

Osserviamo come per tra le nostre variabili, `kend` sia quella che evidenzia una maggior correlazione con le covariate, la correlazione maggiore si ha proprio con `practicePath`.

Capitolo 5

Inferenza

5.1 Stime dei parametri $\hat{\pi}$ delle distribuzioni

Dalle variabili `item1`,..., `item15` e `item1b`,..., `item15b` costruiamo delle nuove variabili indicatrici, di tipo *dummy*. Questa tipologia di variabili assegna valore 1 quando la valutazione dell'osservatore corrisponde alla realtà, facendo riferimento alla variabile `OriginalCode`, ovvero il valore del vetrino assegnato dall'osservatore corrisponde al vero valore riportato nella variabile `OriginalCode`. Per tutti gli altri casi questa variabile assume valore 0, ovvero i casi dove l'osservatore classifica erroneamente il vetrino.
Abbiamo ora una lista di 30 variabili (15 variabili fanno riferimento al tempo 1, ovvero prima che gli osservatori partecipassero al training e le altre 15 sono riferite al tempo due, ovvero dopo aver partecipato al training), di cui i valori saranno tutti 0 o 1. I dati provengono quindi da una distribuzione Bernoulliana. Assunto quindi la distribuzione dei dati di tipo, $Y \sim Be(\pi)$ il nostro obbiettivo è fare inferenza sul parametro π, per far ciò useremo lo stimatore di massima verosimiglianza, la stima della media ($\hat{\pi}$) è la seguente

:

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y} \tag{5.1}$$

Questo stimatore è corretto, consistente ed efficiente, per ulteriori approfondimenti si rimandata a Pace e Salvan (1996). Nella tabella 5.1 sono riportate le stime di $\hat{\pi}_i$, con il relativo intervallo di confidenza ($\alpha = 0.05$) e la deviazione standard. Abbiamo quindi una stima di $\hat{\pi}_i$ per ogni osservatore. La stima di $\hat{\pi}$ totale è 0.46 (con intervallo di confidenza 0.329 - 0.562, deviazione standard pari a 0.497).

	$\hat{\pi}_{pre}$	Est.inf	Est.sup	Dev.std
obs1	0.46	0.34	0.57	0.50
obs2	0.46	0.34	0.57	0.50
obs3	0.81	0.72	0.91	0.39
obs4	0.43	0.31	0.54	0.49
obs5	0.34	0.23	0.45	0.47
obs6	0.43	0.31	0.54	0.49
obs7	0.34	0.23	0.45	0.47
obs8	0.39	0.27	0.50	0.49
obs9	0.34	0.23	0.45	0.47
obs10	0.46	0.34	0.57	0.50
obs11	0.39	0.27	0.50	0.49
obs12	0.47	0.35	0.59	0.50
obs13	0.49	0.37	0.60	0.50
obs14	0.46	0.34	0.57	0.50
obs15	0.43	0.31	0.54	0.49

Tabella 5.1: Stime $\hat{\pi}$ pre training

Da notare nella tabella l'elevata stima di $\hat{\pi}_{obs3} = 0.81$, questo osservatore valuta bene 57 vetrini su 70. I peggiori osservatori sono gli osservatori 9, 8 e 11.

Vediamo ora cosa succede dopo il training nella Tabella 5.2.

	$\hat{\pi}_{post}$	Est.inf	Est.sup	Dev.std
obs1	0.46	0.34	0.57	0.50
obs2	0.60	0.49	0.71	0.49
obs3	0.71	0.61	0.82	0.45
obs4	0.49	0.37	0.60	0.50
obs5	0.47	0.35	0.59	0.50
obs6	0.43	0.31	0.54	0.49
obs7	0.44	0.33	0.56	0.50
obs8	0.57	0.46	0.69	0.49
obs9	0.39	0.27	0.50	0.49
obs10	0.50	0.38	0.62	0.50
obs11	0.44	0.33	0.56	0.50
obs12	0.44	0.33	0.56	0.50
obs13	0.57	0.46	0.69	0.49
obs14	0.40	0.29	0.51	0.49
obs15	0.53	0.41	0.65	0.50

Tabella 5.2: Stime $\hat{\pi}$ post training

La stima totale di $\hat{\pi}$ ora è 0.496 (0.375 - 0.61, deviazione standard 0.499), vi è quindi un leggero aumento della media rispetto a prima. Anche se di poco notiamo un aumento delle valutazioni da parte degli osservatori, soprattutto per gli osservatori 2, 7 e 15. Si nota anche però il peggioramento per alcuni

osservatori, l'osservatore 3 (che resta comunque quello con $\hat{\pi}$ maggiore) e per l'osservatore 14.

5.2 Inferenza sulle stime di $\hat{\pi}$

Useremo ora il Test di McNemar per valutare le distribuzioni delle 15 *dummy* create, si vuole quindi valutare l'effetto del training se influisce o meno sulle distribuzioni dei nostri dati. Il Test di McNemar è dato da:

$$X^2_{MN} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}, \tag{5.2}$$

dove n_{11},n_{21},n_{12} e n_{22} sono ricavate da un tabella 2x2, chiamata anche tabella di contingenza. Viene riportato un esempio nella Tabella 5.3, su due variabili X e Y aventi entrambe due modalità, per approfondimenti si rimanda a Agresti et al. (2000).

	X_1	X_2	totale
Y_1	n_{11}	n_{12}	$n_{1.}$
Y_2	n_{21}	n_{22}	$n_{2.}$
totale	$n_{.1}$	$n_{.2}$	n

Tabella 5.3: Esempio tabella di contingenza

Nella Figura 5.1 viene riportato il grafico del confronto tra le stime. Anche graficamente si nota il leggere miglioramento dopo il training tranne in alcuni casi, sicuramente dove si nota maggior efficacia è per l'osservatore 7 e 8, ma in alcuni casi addirittura la stima pare peggiorare.

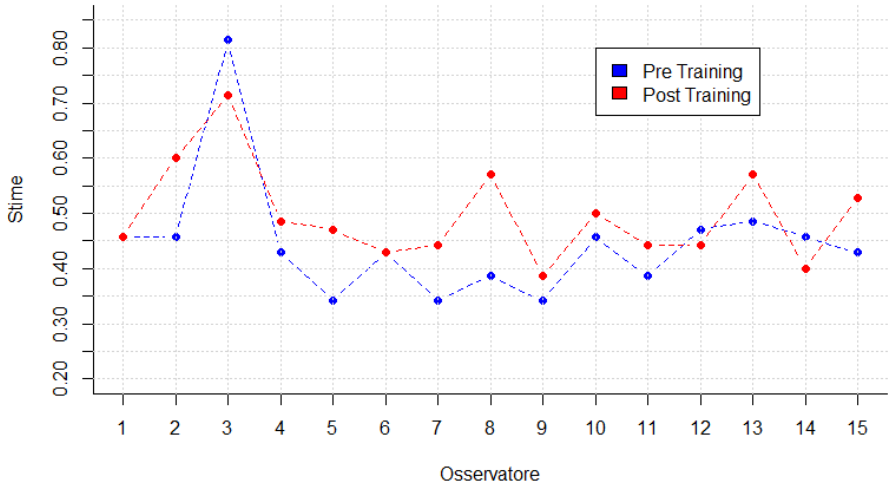


Figura 5.1: Confronto stime di $\hat{\pi}$ pre e post

Applichiamo quindi il test a tutti i nostri i 15 osservatori per valutare se avviene un cambiamento nella classificazione (quindi c'è un effetto del training) oppure no, rimangono invariate (il training non scaturisce nessun effetto). I risultati dei testi tra la distribuzione prima e dopo è riportata in Tabella 5.4.

Oss	p_{value}
obs1	0.931
obs2	0.071
obs3	0.04
obs4	0.571
obs5	0.152
obs6	0.96
obs7	0.219
obs8	0.021
obs9	0.678
obs10	0.689
obs11	0.579
obs12	0.844
obs13	0.231
obs14	0.501
obs15	0.194

Tabella 5.4: Tabella p_{value} dei test di McNemar

Per la maggior parte degli osservatori il test di McNemar accetta l'ipotesi di uguaglianza delle distribuzioni. Fanno eccezione l'osservatore 3 ($p_{value} = 0.04$) e l'osservatore 8 ($p_{value} = 0.021$). Tuttavia abbiamo notato che per l'osservatore 3, il valore di $\hat{\pi}$ peggiora, sembra quindi che il training in questo caso comporti un effetto contrario a quello desiderato. Mentre per l'osservatore 8, la stima di $\hat{\pi}$ migliora, da 0.39 a 0.57. Dunque il training fornisce l'effetto desiderato solamente su un singolo osservatore su 15. Per tutti gli altri osservatori non c'è un cambiamento significativo della distribuzione pre e post training. Il test accetta le ipotesi di uguaglianza delle distribuzioni

per 13 osservatori su 15, per l'osservatore 3 abbiamo notato un peggioramento e solo per l'osservatore 8 possiamo notare un miglioramento significativo dovuto al training.

Conclusioni

Dopo una cenno alla patologia, alla ricerca e alla diagnosi nel Capitolo 1 è stato introdotto il seguente studio, le modalità con cui è stato condotto e le finalità. Nel capitolo 2 sono stati presentati i dati che sono stati usati a tale scopo, analizzando le variabili, la loro natura, introducendo quelle che saranno poi state più rilevanti per il nostro studio. Nel Capitolo 3 si è fatto un breve richiamo teorico alle misure che saranno utilizzate per valutare i nostri dati tra cui il Kappa di Fleiss, la correlazione di Spearman,la correlazione *tau* di Kendall, il Kappa di Cohen classico e la sua versione pesata. Procedendo nel Capitolo 4 si è passati a un'analisi esplorativa di tutte le variabili, riportandone le distribuzioni e le principali statistiche di sintesi. Si sono calcolati gli indici di *agreemnt* sulle variabili più importanti e si sono valutate le differenze principali tra questi indici prima e dopo l'evento oggetto di valutazione. Sono state studiate quali covariate influenzano o meno gli indici di *agreement*. Alla fine nel Capitolo 5, si è cercato di dare una risposta al problema utilizzando alcuni test non parametrici. Quindi l'obbiettivo principale di questo lavoro era analizzare e studiare il possibile grado di accordo nella diagnosi del Papilloma Virus in funzione di un training svolto da alcuni esperti citopatologi incaricati di valutare il virus. Si giunti quindi alle seguenti conclusioni:

- Se l'obbiettivo era quello di dire se il training era efficace o meno: la

risposta è no, il training non produce l'effetto sperato, o meglio, produce un effetto sugli osservatori a cui è stato sottoposto, ma non possiamo dire che questo sia un effetto significativo in termini statistici.

- Gli osservatori hanno comunque valutato discretamente i vetrini, certamente il training non ha migliorato di molto le loro competenze e abilità in questo campo, ma già da prima di questo studio lo loro conoscenze erano a un buon livello.

- Gli osservatori hanno avuto più facilità a classificare giustamente un vetrino con una scarsa presenza di virus rispetto a un vetrino dove la concentrazione dell'agente patogeno era maggiore.

- Analizzando alcune delle variabili covariate, non sono emersi risultati evidenti per quanto riguarda l'esperienza e gli anni di lavoro nel campo per ogni singolo valutatore. Non sempre gli osservatori più anziani, più esperti in citopatologia e patologia hanno dato i risultati migliori. Tali risultati potrebbero essere attribuiti quindi ad attitudini personali, interesse verso il virus e formazione, sarebbe interessate poter disporre di più variabili per poter trovare cosa caratterizzi un "buon osservatore" rispetto agli altri.

- Tuttavia un limite del seguente studio nasce dalle modalità in cui è stato condotto l'osservazione e la valutazione dei vetrini. Molti osservatori hanno riportato di aver avuto difficoltà nella valutazione di vetrini e di aver riscontrato problematiche sulle scannerizzazioni dei vetrini che dovevano valutare, questo potrebbe aver distorto i risultati ottenuti.

5.3 Possibili sviluppi futuri

Il seguente caso di studio dovrebbe essere tenuto come esempio per possibili studi simili. Ripetere questo studio, dando a tutti gli osservatori sotto esame gli stessi materiali e strumenti per una corretta valutazione, potrebbe aiutare a ottenere dei risultati migliori e più affidabili, per far si che non si verifichino come in questo caso dei problemi legati alla qualità della scannerizzazione e nitidezza e colorazione delle immagini, soggette poi a studio e valutazione da parte degli osservatori incaricati. Inoltre, un campione più numeroso di osservatori potrebbe migliorare la qualità e la veridicità dello studio. Infine visti i risultati di questo caso di studio, andrebbe rivisto anche il training. Formare dei professionisti è un costo non da poco, tale investimento andrebbe poi a ripagarsi con la formazione ed esperienza che acquisiscono i soggetti che ne sono sottoposti. Sottoporsi a un training che statisticamente non comporta nessun miglioramento significativo nelle conoscenze e abilità degli osservatori rappresenta solamente un costo e un attività in più per il mondo della sanità. A maggior ragione in un settore così specifico e importante come quello della citopatologia diagnostica, dove sbagliare anche di poco una valutazione, può comportare gravi conseguenze sulla salute delle persone è pesare molto sul sistema sanitario, si pensi al caso dove cure e terapie vengono investite per curare una persona sana o ancora peggio quando una persona malata viene dichiarata in salute o senza nessuna particolare patologia.

Bibliografia

Agresti, Alan, James G Booth*, James P Hobert* e Brian Caffo* (2000). "Random-effects modeling of categorical response data". *Sociological methodology* 30(1), 27–80.

Cronbach, Lee J (1951). "Coefficient alpha and the internal structure of tests". *Psychometrika* 16(3), 297–334.

Doğan, Nurettin Özgür (2018). "Bland-altman analysis: a paradigm to understand correlation and agreement". *Turkish journal of emergency medicine* 18(4), 139–141.

Harris, Tammy e James W Hardin (2013). "Exact wilcoxon signed-rank and wilcoxon mann–whitney ranksum tests". *The stata journal* 13(2), 337–343.

Lands, JR e GG Koch (1977). "The measurement of observer agreement for categorial data". *Biometrics* 33(1), 159–74.

Lodi, G, M Tarozzi, E Baruzzi, D Costa, R Franchini, F D'amore, A Carrassi e N Lombardi (2021). "Epidemiology and risk factors". EDRA.

Mitani, Aya A, Phoebe E Freer e Kerrie P Nelson (2017). "Summary measures of agreement and association between many raters' ordinal classifications". *Annals of Epidemiology* 27, 677–685.

Muliere, Pietro (1976). "Una nota sul coefficiente di correlazione tra l'indice g di cograduazione di gini e l'indice di kendall". *Giornale degli economisti e annali di economia* 35(9/10), 627–633.

Nelson, Kerrie P e Don Edwards (2008). "On population-based measures of agreement for binary classifications". *Canadian journal of statistics* 36(3), 411–426.

Pace, Luigi e Alessandra Salvan (1996). *Introduzione alla statistica: inferenza, verosimiglianza, modelli.-2001.-xvi, 422 p.* Cedam.

Solomon, Diane, Diane Davey, Robert Kurman, Ann Moriarty, Dennis O'Connor, Marianne Prey, Stephen Raab, Mark Sherman, David Wilbur, Thomas Wright Jr, Nancy Young, for the Forum Group Members e the Bethesda 2001 Workshop (2002). "The 2001 Bethesda SystemTerminology for Reporting Results of Cervical Cytology". *Jama* 287(16), 2114–2119.

Tan, Siang Yong e Yvonne Tatsumura (2015). "George papanicolaou (1883–1962): discoverer of the pap smear". *Singapore medical journal* 56(10), 586.

Sitografia

Lega Italiana Lotta Ai Tumori, LILT (2019). *Papilloma virus*. https://www.legatumori.mi.it/fai-prevenzione/papilloma-virus/. Online, ultimo accesso 08 giugno 2022.