Rappresentazione grafica degli item a tre categorie: un approccio visivo all'analisi delle preferenze

Visual Representation of Trichotomous Items: A Graphical Approach to Analyzing Preferences

Carlo Di Pietrantonj[1]

[1] S.S.D. Epidemiologia, promozione della salute e coordinamento attività di prevenzione.
email: cdipietrantonj@aslcn2.it

Riassunto: La visualizzazione dei dati è uno strumento fondamentale nell'analisi statistica, poiché consente di evidenziare strutture e schemi nascosti. In particolare quando vengono utilizzate batterie di item categorici a tre modalità come "sì / no / non so", come ad esempio in questionari di opinione o in scale di valutazione e indagini percettive, la rappresentazione tabellare non sempre risulta immediata da leggere, mentre un grafico aiuta a cogliere con immediatezza differenze e similarità. Questo lavoro propone un approccio eseguibile tramite un foglio di calcolo per rappresentare su di un grafico gli item a tre categorie. Ogni item viene collocato in un piano cartesiano in funzione delle proporzioni delle risposte positive (SI) e negative (NO) e neutre (Non So), il risultato è una mappa visiva che consente di individuare la risposta prevalente di ciascun item. Il lavoro presenta due tipologie di grafico triangolare e propone una applicazione grafica del test del chi-quadrato per delimitare le aree corrispondenti a distribuzioni eterogenee delle risposte interne ad ogni item.

Abstract: Data visualization is a fundamental tool in statistical analysis because it allows hidden structures and patterns to be revealed. Particularly when sets of three-mode categorical items such as "yes / no / don't know" are used, such as in opinion questionnaires or in rating scales and perceptual surveys, tabular representation is not always straightforward to read, whereas a graph helps to capture differences and similarities with immediacy. This paper proposes a spreadsheet-executable approach to represent three-category items on a graph. Each item is placed in a Cartesian plane according to the proportions of positive (YES) and negative (NO) and neutral (Don't Know) responses; the result is a visual map that allows one to identify the prevail-

Bollettino Epidemiologico e Prospettive di Sanità Pubblica (BE.PSP), anno 2025
DOI: 10.14672/bepsp2025272-292

ing response of each item. The paper presents two types of triangular graph and proposes a graphical application of the chi-square test to delineate areas corresponding to heterogeneous distributions of responses within each item.

Introduzione

La data visualization rappresenta una componente essenziale dell'analisi dei dati, finalizzata a mettere in luce strutture che difficilmente emergerebbero attraverso riepiloghi tabellari o sintetici [*Friendly 2000, Midway 2020*]. Una rappresentazione visiva accattivante e ben progettata agevola la comprensione e l'esplorazione dei dati [*Tukey 1977, Midway 2020*] e stimola riflessione critica e consente di comunicare efficacemente anche informazioni complesse. Secondo altri ricercatori le rappresentazioni grafiche possono assolvere funzioni analitiche, orientate all'esplorazione e alla modellizzazione, oppure comunicative, finalizzate a informare o persuadere [*Friendly 2000, Midway 2020*]. In entrambi i casi la componente estetica, che non può prescindere da quella funzionale, deve contribuire a evidenziare il messaggio statistico.

In questo lavoro presenteremo due metodi per rendere graficamente gli item che si presentano come variabili con tre categorie, mutuamente esclusive, ad es. (Si, No, Non-So/Non-Risponde) oppure (Gradito, Sgradito, Indifferente/Non-Risponde). Il primo metodo utilizza un piano cartesiano per rappresentare sull'asse orizzontale la differenza fra la proporzione di risponde "Si" / "D'accordo" e "No" / "In disaccordo", e sull'asse verticale la proporzione di risposte "non so" / "indifferente", in questo grafico ognuna delle possibili distribuzioni delle risposte all'interno di un item definisce una posizione sul grafico, l'insieme di tutte le posizioni forma una figura a triangolo isoscele; mentre il secondo grafico noto come Trilinear o Trinomial plot o Ternary diagram [*Friendly 2000, Fraser 2017*] **è** uno scatterplot concepito con un sistema di coordinate diverso dal precedente che permette di visualizzare le proporzioni di risposte in un diagramma a triangolo equilatero. Entrambi i grafici permettono di visualizzare in un piano le distanze reciproche tra le distribuzioni di risposta dei vari item, tuttavia nessuno dei due grafici mette in evidenza per quali item le proporzioni di risposta risultano statisticamente eterogenee. Obiettivo del lavoro **è** confrontare questi due tipi di rappresentazione grafica e proporre un metodo, applicabile con un semplice foglio di calcolo, per suddividere il grafico in zone che permettano di individuare quegli item le cui distribuzioni delle riposte risultano eterogene. Le dimostrazioni dei risultati matematici sono esposte in appendice e non sono necessarie per la comprensione operativa dei diagrammi.

La Struttura dei Dati

In generale la struttura dei dati è una tabella con K righe e tre colonne, le righe (ovvero gli item) sono gli oggetti della rappresentazione (le unità di rappresentazione) e le tre colonne sono i punti della scala, categorie mutuamente esclusive; inoltre assumiamo che tutti gli N soggetti (le unità statistiche) abbiano dato una delle tre possibili risposte per ognuno dei K item. Quindi, leggendo la tabella lungo le righe vediamo la ripartizione dei soggetti per le tre possibili riposte date allo specifico item (tabella 1); ad es. per l'i-esimo-item (item-i), $F_{i,a}$ rappresenta il numero di individui che ha dato la Risposta(a) e $P_{i,a}$ la corrispondente proporzione sul totale *N, mentre* $F_{i,b}$ è il numero di individui che ha dato la Risposta(b) e $P_{i,b}$ la corrispondente proporzione sul totale *N,* infine $F_{i,c}$ è il numero di individui che ha dato la Risposta(c), e $P_{i,c}$ la relativa proporzioni sul totale N. Dove $F_{i,a} + F_{i,b} + F_{i,c} = N$ e quindi $P_{i,a} + P_{i,b} + P_{i,c} = 1$.

Per rappresentare su di grafico bidimensionale ciascun item in base alla distribuzione delle risposte tra le sue categorie, possiamo utilizzare due sistemi di coordinate che producono due differenti grafici, il primo che chiameremo "grafico Triangolare" e il secondo "diagramma Trilineare".

Tabella 1: Struttura dei dati – Frequenze delle Risposte

	%Risposte (a)	%Risposte (b)	%Risposte (c)	Totale rispondenti
Item-1	$P_{1,a} = F_{1,a}/N$	$P_{1,b} = F_{1,b}/N$	$P_{1,c} = F_{1,c}/N$	N
Item-2	$P_{2,a} = F_{2,a}/N$	$P_{2,b} = F_{2,b}/N$	$P_{2,c} = F_{2,c}/N$	N
⋮	⋮	⋮	⋮	⋮
Item-i	$P_{i,a} = F_{i,a}/N$	$P_{i,b} = F_{i,b}/N$	$P_{i,c} = F_{i,c}/N$	N
⋮	⋮	⋮	⋮	⋮
Item-K	$P_{k,a} = F_{k,a}/N$	$P_{k,b} = F_{k,b}/N$	$P_{k,c} = F_{k,c}/N$	N

Le coordinate del grafico Triangolare sono:

$$(1) \quad [X = P_a - P_b;\ Y = P_c];$$

le quali posizionano ogni possibile item all'interno di un grafico a triangolo isoscele dove sull'asse orizzontale viene rappresentata la differenza delle proporzioni tra le due categorie principali, ad esempio "sì" vs "no" o "gradito" vs "sgradito", mentre sull'asse verticale viene rappresentata la proporzione di risposte appartenenti alla terza categoria, come "non so" o "indifferente", che indica una componente informativa di incertezza o neutralità. In questa rappresentazione gli item che si collocano a destra presentano una prevalenza di risposte nella categoria (a) ad es. "sì"; viceversa se si collocano

a sinistra è prevalente la categoria (b) ad es. "no", infine si collocano in alto quegli item con prevalenza di risposte (c) ad es. "non so".

Mentre, le coordinate del diagramma Trilineare [*Friendly 2000, Fraser 2017*] sono:

$$(2) \quad \left[X = P_{i,a} + \frac{P_{i,c}}{2};\ Y = P_{i,c}\sqrt{3}/2\right]$$

che posizionano gli item su una figura a triangolo equilatero nel quale ogni vertice rappresenta una delle tre categorie. Quindi, tanto più un item si colloca vicino a un vertice (ad es (a)) tanto prevalente risulterà quella categoria per quell'item, mentre tanto più un item si colloca vicino al lato opposto dei quel vertice tanto più importanti saranno le altre due categorie per quell'item.

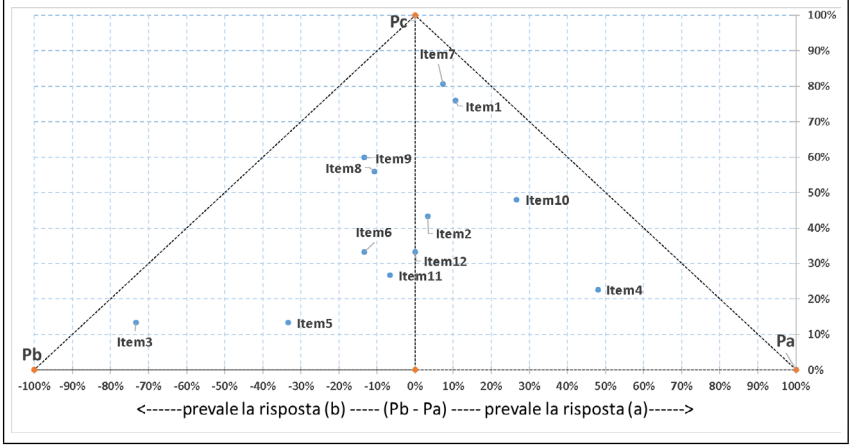
Per illustrare la costruzione e l'interpretazione dei due grafici consideriamo i dati In tabella 2 dove mostriamo un esempio con dati simulati, di una indagine su di un campione 150 soggetti ai quali è stato chiesto di esprimere la propria preferenza o opinione su 12 possibili temi utilizzando le risposte Si / No / Non So; come ad esempio "ritieni che le attuali politiche dell'UE in tema di protezione ambientale siano efficaci? Oppure "sei favorevole a estendere il diritto di voto alle persone di 16 e 17 anni?".

Nella tabella 2 si sono riportate per ogni item le frequenze assolute (a, b, c) e le frequenze relative (Pa, Pb, Pc), espresse in forma percentuale, delle risposte dei 150 soggetti, infine sono riportate i due tipi di coordinate (in forma percentuale), che possono essere disegnate utilizzando qualsiasi software che permette di realizzare grafici a diagramma cartesiano (scatterplot).

In figura 1 sono rappresentati le posizioni dei 12 item secondo le coordinate del grafico Triangolare, il quale è concepito per enfatizzare il contrasto tra la frequenza relativa delle risposte (a) e la frequenza relativa delle risposte (b), rappresentando sull'asse orizzontale la differenza delle proporzioni (Pa-Pb) e simultaneamente visualizzare il peso delle risposte (c) rappresentate sull'asse verticale.

Tabella 2: dati simulati di risposte a 12 item.										
	SI	No	Non so	Tot.	Pa	Pb	Coordinate grafico Triangolare		Coordinate diagramma Trilineare	
	(a)	(b)	(c)	N	x 100	x 100	X = Pa–Pb	Y = Pc	X = Pa+Pc/0.5	Y = Pc√3/2
							x 100	x 100	x 100	x 100
Item1	26	10	114	150	17.3%	6.7%	10.7%	76.0%	55.3%	65.8%
Item2	45	40	65	150	30.0%	26.7%	3.3%	43.3%	51.7%	37.5%
Item3	10	120	20	150	6.7%	80.0%	-73.3%	13.3%	13.3%	11.5%
Item4	94	22	34	150	62.7%	14.7%	48.0%	22.7%	74.0%	19.6%
Item5	40	90	20	150	26.7%	60.0%	-33.3%	13.3%	33.3%	11.5%
Item6	40	60	50	150	26.7%	40.0%	-13.3%	33.3%	43.3%	28.9%
Item7	20	9	121	150	13.3%	6.0%	7.3%	80.7%	53.7%	69.9%
Item8	25	41	84	150	16.7%	27.3%	-10.7%	56.0%	44.7%	48.5%
Item9	20	40	90	150	13.3%	26.7%	-13.3%	60.0%	43.3%	52.0%
Item10	59	19	72	150	39.3%	12.7%	26.7%	48.0%	63.3%	41.6%
Item11	50	60	40	150	33.3%	40.0%	-6.7%	26.7%	46.7%	23.1%
Item12	50	50	50	150	33.3%	33.3%	0.0%	33.3%	50.0%	28.9%

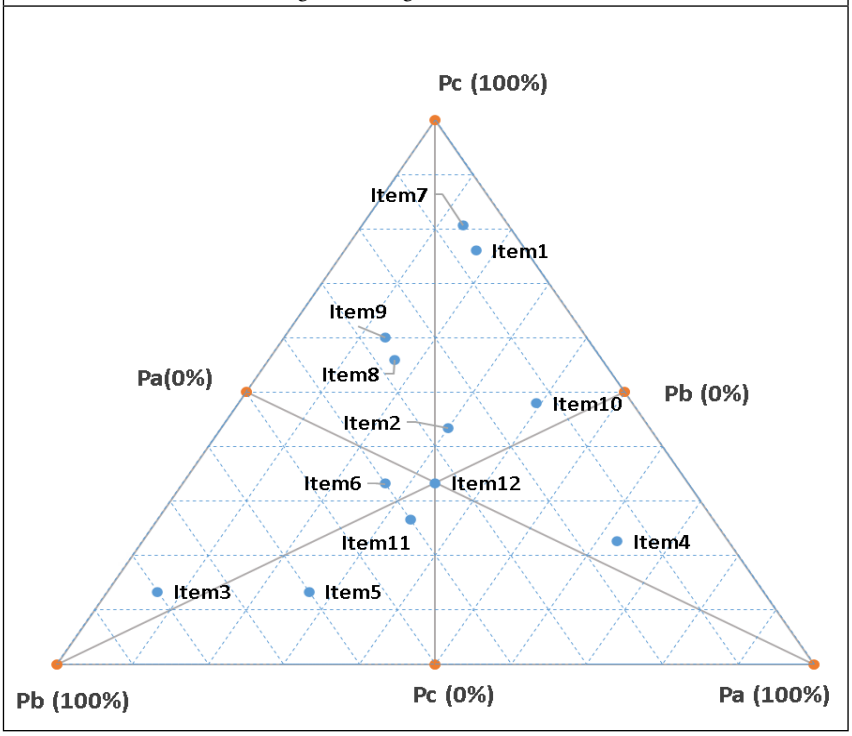
Figura 1: Grafico Triangolare



La rappresentazione fornita In figura 1 consente una lettura delle differenti distribuzioni delle risposte per ciascun item. La distanza dei punti dall'asse verticale al centro del triangolo fornisce informazioni quali/quantitative su quanto sia marcato il contrasto tra le due categorie principali (a) vs (b), mentre la distanza dalla base informa su quanto sia rilevante la quota di risposte (c). Ad esempio gli Item 1, e 7 sono caratterizzati da un'alta preva-

lenza (>70%) di risposte (c), mentre per gli item 3 e 5, prevale la risposta (b) e per l'item 4 la risposta (a), per gli item 8, e 9 prevalgono le risposte (c) e (b), mentre per l'item 10 prevalgono le risposte (c) e (a), infine l'item 2, 6, 11 e 12 si trovano raggruppati nel centro del grafico. L'item 12 in particolare mostra avere la percentuale delle risposte (a) eguale alla percentuale delle risposte (b) e si colloca sull'asse centrale alla quota pari al 33,3%. La rappresentazione dei dati in tabella 2 secondo il diagramma Trilineare è mostrata In figura 2.

Figura 2: Diagramma Trilineare



Questo diagramma è concepito per rappresentare simultaneamente le tre percentuali su un triangolo equilatero; tuttavia la corretta lettura delle posizioni dei punti non avviene lungo due assi ortogonali come in un diagramma cartesiano, ma lungo le tre bisettrici del triangolo equilatero. Il punto d'incontro delle tre bisettrici è la posizione occupata dagli item che presentano distribuzione uniforme tra le categorie. Tanto più un item si avvicina ad un vertice del triangolo tanto più alta sarà la percentuale di risposte per la categoria corrispondente, mentre tanto più si avvicina al lato opposto (percorrendo la bisettrice) tanto minore sarà la percentuale di risposte per quella categoria, e tanto maggiore sarà il peso percentuale assunto delle altre due categorie. Quindi muovendosi lungo le bisettrici verso il vertice, vedremo

aumentare la percentuale per quella categoria, viceversa muovendoci verso la base la vedremo diminuire; mentre le percentuali delle altre due categorie rimarranno in equilibrio. Consideriamo ad esempio l'item 3, il quale si colloca poco sopra la prima linea orizzontale a partire dalla base, il lato Pc(0%), che corrisponde a Pc circa uguale a 10%, simultaneamente si colloca sulla ottava linea diagonale a partire dal lato Pb(0%) che corrisponde a Pb = 80%, infine sempre simultaneamente si colloca di poco al disotto della prima linea diagonale a partire dal lato Pa(0%) che corrisponde ad un valore di poco inferiore al 10%. In generale la lettura dei posizionamenti di tutti gli item conduce ai medesimi raggruppamenti letti nel grafico triangolare illustrato precedentemente.

Entrambe le rappresentazioni sono realizzabili tramite un qualsiasi foglio di calcolo che preveda fra i possibili grafici i diagrammi a dispersione (scatterplot), inoltre sono descrittivamente sovrapponibili. Tuttavia, la lettura del grafico Triangolare (figura 1) si appoggia direttamente sul piano cartesiano; mentre la lettura del diagramma Trilineare (figura 2) risulta meno immediata, poiché deve avvenire considerando come assi di riferimento, non gli assi cartesiani, ma le tre bisettrici del triangolo, inoltre la costruzione delle linee di riferimento, necessarie alla sua lettura, risulta particolarmente laboriosa. Sottolineiamo che il grafico Triangolare rappresenta direttamente il contrasto tra due particolari categorie (a) vs (b) relativamente alla terza (c), quindi risulta adeguato nei casi in cui tale confronto è interessante per il processo decisionale, ad esempio quando è importante individuare se la frazione dei "d'accordo" è prevalente rispetto ai "disaccordo" o viceversa, condizionatamente a chi preferisce una risposta "neutra". Mentre il diagramma Trilineare è un diagramma composizionale che non accorda preferenze particolari alle categorie. Infine entrambe le rappresentazioni permettono di stabilire quali item presentano al loro interno distribuzioni di risposte particolarmente disomogenee solo da un punto di vista qualitativo. Riuscire a visualizzare sui grafici quelle posizioni che riflettono differenze statisticamente significative nella distribuzione delle risposte, renderebbe le rappresentazioni grafiche non solo descrittive ma anche analitiche.

Test di significatività

Quando si desidera analizzare la distribuzione delle risposte tra categorie mutuamente esclusive, la configurazione di riferimento usualmente considerata è quella della distribuzione uniforme, l'equa ripartizione dei rispondenti tra le tre categorie. In altre parole desideriamo identificare gli item la cui distribuzione interna delle proporzioni di risposta si discosta in modo statisticamente significativo dalla distribuzione (1/3; 1/3 1/3) ovvero (33.3%, 33.3%, 33.3%). Nei due grafici utilizzati, la posizione corri-

spondente alla distribuzione uniforme è rappresentata dal punto "item12": il quale nel grafico Triangolare ha coordinate [0; 1/3] ovvero [0; 33,3%], mentre nel diagramma Trilineare corrisponde al centro del triangolo, con coordinate ovvero [50,0%; 28,9%]. L'obiettivo è definire, all'interno di ciascuna rappresentazione grafica, un'area di omogeneità: ossia una regione in cui i punti rappresentano distribuzioni non significativamente diverse dalla distribuzione uniforme (ad esempio al livello del 5%). A questo fine è possibile utilizzare l'usuale test del chi-quadrato (Chi2) [Siegel 1998] la cui l'usuale formula:

$$(3) \qquad test = \sum_{i} \frac{(Osservati - Attesi)^2}{Attesi}$$

che applicata alle frequenze relative può esser riscritta

$$(4) \qquad test = \sum_{i} \frac{\left(N \cdot \frac{Osservati}{N} - \mathrm{N} \cdot \frac{Attesi}{N}\right)^2}{N \cdot \frac{Attesi}{N}} = \sum_{i} \frac{(N \cdot P_{Oss} - \mathrm{N} \cdot P_{Att})^2}{\mathrm{N} \cdot P_{Att}}$$

Quindi per di ciascuno degli item (i) la (3) assumerà la forma:

$$(5) \qquad test = N\left(\frac{(P_{i,a} - \mathrm{E}_{i,a})^2}{\mathrm{E}_{i,a}} + \frac{(P_{i,b} - \mathrm{E}_{i,b})^2}{\mathrm{E}_{i,b}} + \frac{(P_{i,c} - \mathrm{E}_{i,c})^2}{\mathrm{E}_{i,c}}\right)$$

La (5) si distribuirà come una distribuzione Chi2 con 2 gradi di libertà, pertanto volendo definire l'insieme dei punti del grafico che rendono statisticamente significativo il test (5) deve essere soddisfatta la diseguaglianza: $test > d$; dove $\left(\mathrm{E}_{i,a} = \frac{1}{3};\ \mathrm{E}_{i,b} = \frac{1}{3};\ \mathrm{E}_{i,c} = \frac{1}{3}\right)$; mentre d rappresenta il valore soglia per un test del chi-quadrato con due gradi di libertà al livello α%; ad esempio per un singolo test: $d = 5.99$ *per* $a = 5\%$, mentre N è la dimensione campionaria. Riorganizzando la (5) risulta facile dimostrare che nel caso del grafico triangolare avremo un'ellisse con semi assi di grandezza proporzionale al rapporto *d/N* mentre per il diagramma Trilineare si giunge ad un cerchio il cui raggio sarà sempre proporzionale al rapporto *d/N* (appendice dim1 e appendice dim2). Tuttavia la sovrapposizione dell'ellisse sul grafico Triangolare, come la sovrapposizione del cerchio sul diagramma Trilineare, potrebbe risultare laboriosa utilizzando un normale foglio di calcolo, inoltre nel caso del grafico Triangolare concepito per visualizzare il contrasto tra *(a) e (b)*, condizionatamente alla frequenza di riposte *(c)*; il test (5) che è un test chi2-globale non coglierebbe questo aspetto che risulterebbe meglio analizzato da un test sulla differenza delle proporzioni.

Test sulla differenza di proporzioni – area di omogeneità parabolica

Possiamo costruire il test statistico che individui l'area del grafico triangolare le cui posizioni corrispondono agli item per le quali la differenza tra la

proporzione delle risposte (a) e le risposte (b) risulti statisticamente significativa. A questo scopo è possibile utilizzare il test di Wald [Piccolo1996, Grassi1994].

$$(6) \qquad \chi^2_{(df=1)} \approx \frac{X^2}{Var(\mathrm{X})} > d$$

Dove *d* è il valore soglia del test del Chi-quadrato con un grado di libertà al livello α%; mentre e . Quindi ricordando che *Y=Pc* risulta facile da dimostrare (appendice dim3) che è la (6) può essere riscritta come

$$(7) \qquad \frac{NX_i^2}{[(1-Y_i)-X_i^2]} \geq d$$

Dalla (7) è facile ottenere l'equazione di una parabola:

$$(8) \qquad Y_i \geq 1 - \mathrm{X}_i^2 \left(\frac{N+d}{d}\right)$$

La parabola (8) ha il suo massimo nel punto di coordinata [0;100%] e individua al suo esterno l'area nella quale le differenze tra le proporzioni delle risposte (a) e (b) non sono statisticamente significative, è facile osservare che l'ampiezza dell'area parabolica è inversamente proporzionale ad N, tanto maggiore sarà la dimensione campionaria tanto minore sarà l'area di omogeneità.

Correzione per confronti multipli

L'applicazione ripetuta di test statistici sul medesimo campione, con più sottogruppi o più di variabili, aumenta la probabilità dichiarare statisticamente significative differenze in realtà dovute al caso; ciò è noto come inflazione dell'errore di I tipo. Per limitare questo rischio si applicano delle correzioni che, all'aumentare del numero dei test, rendono più stringente la soglia di significatività. Tra i diversi approcci esistenti quello di Bonferroni è il più semplice e consiste nel dividere il livello di significatività α%, per il numero dei test effettivamente eseguiti. Nel nostro caso il nuovo livello di significatività sarà ottenuto dividendo α per il numero dei test eseguiti ovvero il numero deli item che vengono confrontati. Più formalmente: il metodo proposto che confronta k item di fatto esegue k test statistici, pertanto il valore critico *d*, che per i test basati sulla distribuzione chi2 può essere calcolato tramite la funzione presente nei fogli di calcolo: *d = inv.chi(a/k; df)* dove *a = livello di significatività* (tipicamente 5%), *k = il numero dei test* (ovvero il numero degli item), *df = gradi di libertà.*

Quindi, nel caso con 12 item e quindi 12 test al 5%, il nuovo livello di significatività sarà pari a 0,05/12 = 0,00417. Quindi, per il Chi2-

globale (5) *d=inv.chi(0,00417;2) = 10.961* e per il test di Wald (8) *d=inv.chi(0,00417;1)= 8,210*.

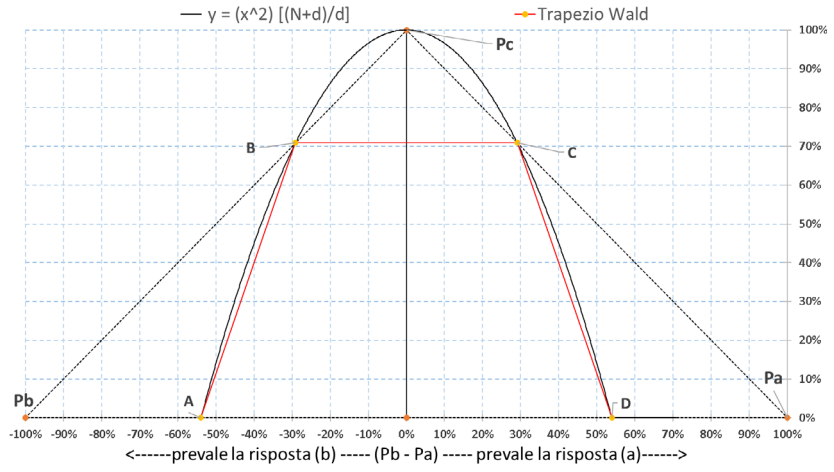
Approssimazione dell'area parabolica

Tracciare una parabola con un foglio di calcolo può risultare laborioso, tuttavia è possibile semplificare il problema utilizzando solo i 4 punti di intersezione della (8) o della (9) con i lati del grafico triangolare (si veda appendice dim.3). In tabella 3 sono riportati i punti di intersezione con la base del triangolo (Y=0) (punti A e D), e i punti di intersezione con i lati del triangolo (punti B e C).

Tabella 3: coordinate del trapezio	
il test di Wald (8)	
[X; Y]	**[X; Y]**
B: $\left[-\frac{d}{N+d}; \frac{N}{N+d}\right]$	**C:** $\left[\frac{d}{N+d}; \frac{N}{N+d}\right]$
A: $\left[-\sqrt{\frac{d}{N+d}}; 0\right]$	**D:** $\left[\sqrt{\frac{d}{N+d}}; 0\right]$

Questi punti definiscono una zona trapezoidale interna alla parabola (figura 3) che individua al suo interno tutte le posizioni la cui differenza tra le proporzioni delle risposte (a) e (b) non risulta statisticamente significativa. Bisogna osservare che il trapezio cerca di approssimare gli archi della parabola, quindi possono esserci posizioni sul grafico Triangolare esterne al trapezio, ma interne alla parabola, pertanto, il giudizio basato sull'area trapezoidale potrebbe produrre falsi positivi (figura 3).

Figura 3: Grafico Triangolare – es. area di omogeneità parabolica e trapezoidale



Esempio pratico1: area di omogeneità parabolica

Riprendiamo in tabella 4 l'esempio numerico illustrato in tabella 2 limitatamente al grafico Triangolare poiché siamo interessati a evidenziare la significatività statistica della differenza tra le proporzioni (a) e (b).

Tabella 4: Esempio di applicazione grafico Triangolare

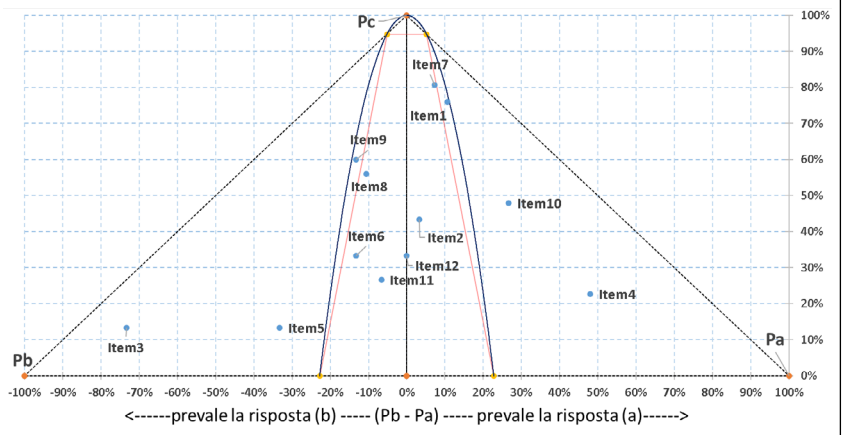
	Coordinate Grafico Triangolare		Coordinate Punti di Riferimento			Coordinate Trapezio			Test di Wald	
	X = Pa–Pb	Y = Pc	Etich.	X	Y	Etich.	X	Y	valore	Sign.
Item1	10.7%	76.0%	Pb	-100%	0	A	-22.8%	0.0%	7.465	
Item2	3.3%	43.3%	Pa	100%	0	B	-5.2%	94.8%	0.295	
Item3	-73.3%	13.3%	Pc	0	100%	C	5.2%	94.8%	245.27	*
Item4	48.0%	22.7%		-100%	0	D	22.8%	0.0%	63.654	*
Item5	-33.3%	13.3%		0	0				22.059	*
Item6	-13.3%	33.3%		0	100%				4.11	
Item7	7.3%	80.7%							4.292	
Item8	-10.7%	56.0%							3.982	
Item9	-13.3%	60.0%							6.977	
Item10	26.7%	48.0%							23.762	*
Item11	-6.7%	26.7%							0.915	
Item12	0.0%	33.3%							0	

Sign.: Indica con un asterisco (*) quando il valore del test è superiore al valore critico *d=8,210*

Ricordando che il numero dei rispondenti è N=150, che il valore critico del test d deve essere calcolato con la correzione di Bonferroni considerando 12 confronti per il test (6) con df=1, quindi d = inv.chi(0,00417;1) = 8,210. In tabella 4 sono riportate sia le tre serie dati necessarie a costruire con un foglio di calcolo il grafico triangolare e l'area di omogeneità trapezoidale; inoltre in tabella 5 sono riportati il valore del test di Wald per ogni item calcolato dalla (7).

Le colonne denominate "Coordinate Punti di Riferimento" aggiunte allo scatterplot e unite con una linea permettono di rappresentare sul diagramma la forma triangolare. Da notare che questa sequenza non è l'unica possibile, tuttavia l'ordine è importante per permettere al foglio di calcolo di disegnare il triangolo congiungendo i punti tramite una linea, inoltre le tre coordinate senza etichette consentono di completare il triangolo disegnando, in questo caso, il lato (PcPb) e l'altezza del triangolo dal punto medio della base [0,0] al vertice (Pc). Mentre le colonne denominate "Coordinate Punti Trapezio", ottenute dalle formule in tabella 3, inserite nello scatterplot come terza serie e unite con una linea permettono di disegnare la forma trapezoidale, anche in questo caso l'ordine è importante. Il risultato della proiezione sul piano cartesiano di queste tre serie di punti è rappresentato in figura 4, dove abbiamo anche disegnato la parabola usando l'equazione (8).

Figura 4: Grafico Triangolare dati tabella 4 - area di omogeneità parabolica e trapezoidale



Dalla lettura della figura 4 emerge che per gli item 3 e 5 prevalgono in modo statisticamente significativo (livello 5% corretto con Bonferroni) le risposte (b), mentre per gli item: 10 e 4 prevalgono in modo statisticamente significativo le risposte (a). Infine gli item 9 e 1 risultano molto vicini ai lati del trapezio pertanto ricadono nell'area compresa tra il lato del trapezio e l'arco di parabola quindi per loro non risulta statisticamente significativa la

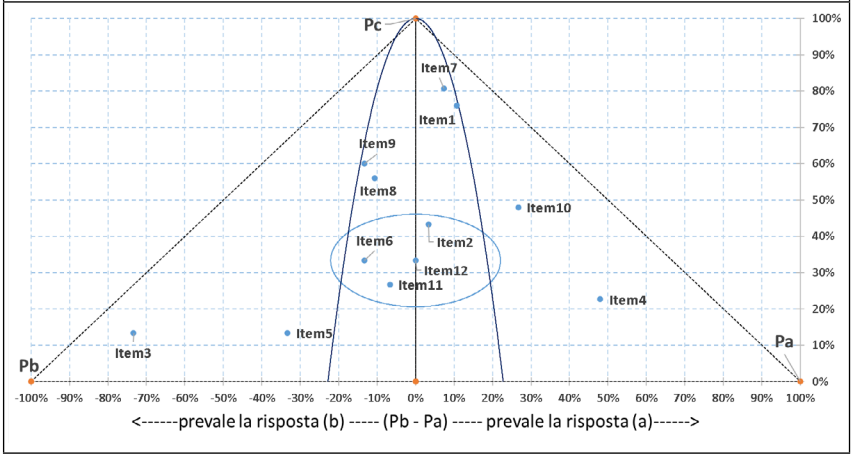
differenza delle proporzioni Pa Pb come è possibile verificare anche dalla tabella 4.

Costruire le coordinate della parabola con un foglio di calcolo e inserirla come serie di punti nello scatterplot potrebbe non essere particolarmente laborioso, tuttavia usare i punti del trapezio semplifica il lavoro di preparazione per la rappresentazione grafica, sebbene obblighi a verificare il risultato dei punti vicini ali lati obliqui controllando il test statistico.

Esempio pratico2: area di omogeneità ellittica e parabolica

Per completezza mostriamo il grafico Triangolare con l'ellisse di omogeneità (test chi2 (5) al livello del 5% con correzione di Bonferroni, quindi d = 10,961) che al suo interno include gli item (2, 6,11 e 12) la cui distribuzione interna delle risposte non è differente in modo statisticamente significativo dalla distribuzione uniforme (33,3%; 33,3%; 33,3%), mentre per gli item: 1,7,8 e 9 è significativa la prevalenza delle risposte (c). l'intersezione dell'ellisse con la parabola evidenzia la possibilità che esistano item con test di Wald (6) statisticamente significativo, ma che non risultano diversi in modo statisticamente significativo dalla distribuzione uniforme (il test globale del chi-quadrato (5)).

Figura 5: Grafico Triangolare dati tabella 4 - area di omogeneità parabolica ed ellittica



Per facilitare l'esecuzione di questa versione del grafico triangolare (figura 5), per chi ha familiarità con il software R, in appendice è possibile utilizzare un programma scritto per ottenere il grafico triangolare che rappresenti simultaneamente l'area di omogeneità ellittica e parabolica.

Conclusioni

Questo lavoro presenta due differenti approcci grafici per la rappresentazione e l'interpretazione di item a tre categorie: il diagramma Trilineare e il grafico Triangolare. Il primo pur presentando il vantaggio di rappresentare simultaneamente la distribuzione delle risposte sulle tre categorie, non appare facilmente eleggibile, inoltre la sua realizzazione tramite un foglio di calcolo si presenta laboriosa. Mentre il grafico triangolare è concepito per mostrare il contrasto tra due categorie principali (es. "sì" vs "no") in relazione alla terza ("non so" o "indifferente"), risulta più immediato da leggere e facile da realizzare tramite un foglio di calcolo.

Il lavoro inoltre presenta un metodo per integrare i test statistici di significatività direttamente nelle rappresentazioni grafiche. L'impiego del test del chi-quadrato e del test di Wald ha permesso di delimitare aree di omogeneità che evidenziano gli item la cui distribuzione interna delle risposte si discosta significativamente dalla distribuzione uniforme fra le categorie (area di omogeneità ellittica) e bilanciata tra le proporzioni delle due categorie principali (area di omogeneità parabolica). La possibilità di rappresentare tali aree direttamente sul grafico arricchisce la dimensione descrittiva con una valenza analitica, pur mantenendo la semplicità di implementazione in un foglio di calcolo.

Il grafico triangolare permette di rappresentare set di item a tre categorie in ambiti come le indagini percettive, i questionari di opinione o le valutazioni soggettive, dove l'immediatezza comunicativa rappresenta un requisito essenziale. Inoltre, rappresenta uno strumento versatile e soprattutto accessibile poiché facilmente realizzabile con un foglio di calcolo anche a chi non ha familiarità con software statistici.

Bibliografia

Friendly 2000: Friendly M, Visualising Categorical Data, Cary NC SAS Institute Inc.2000

Midway 2020: Midway SR. Principles of effective data visualization. Patterns [Internet]. 2020 Dec 11;1(9):100141. Available from: https://doi.org/10.1016/j.patter.2020.100141

Fraser 2017: L. Fraser Jackson and Mohammed S Khaled Plotting labour force status shares: Interdependence and ternary plots. Working papers in economics and finance, published by the School of Economics and Finance 2017

Armitage 1996: Armitage P, Berry G. Statistica Medica – metodi statistici per la ricerca in medicina. McGraw-Hill 1996

Tukey 1977: Tukey J W, Exploratory data analysis, Addison Wesley, Reading MA 1977

Siegel 1998: Siegel S, Castellan N J. Non Parametric Statistics for the Behavioral Sciences. Mcgraw Hill 1998

Zelterman 2004: Daniel Zelterman, Discrete Distributions, Applications in the Health Sciences. Wiley 2004.

Mood 1992: Mood M.A, Graybill F.A. Boes D.C. Introduzione alla statistica. McGraw Hill 1992

Piccolo 1998: Piccolo D. Statistica. Il Mulino 1998

Grassi 1994: Statistica in Medicina un approccio basato sulla verosimiglianza, McGraw Hill 1994

Appendice dim.1: ellisse di omogeneità

A partire dall'usuale test del chi-quadrato

$$(1.1) \qquad N\left(\frac{(P_{i,a}-\frac{1}{3})^2}{1/3}+\frac{(P_{i,b}-\frac{1}{3})^2}{1/3}+\frac{(P_{i,c}-\frac{1}{3})^2}{1/3}\right) > d$$

Ricordando che $P_{i,a}+P_{i,b}+P_{i,c}=1$, la (1.1) può essere semplificata come segue:

$$(1.2) \qquad {P_{i,a}}^2+{P_{i,b}}^2+{P_{i,c}}^2 > \frac{d}{3\,N}+\frac{1}{3}$$

Nel grafico dove $[X = P_a - P_b;\ Y = P_c]$l'insieme dei punti che soddisfano la (1.2); sarà centrato sul punto [0; 1/3]. Poiché: $P_{i,a}^2+P_{i,b}^2=(P_{i,a}-P_{i,b})^2+2P_{i,a}P_{i,b}$ il quadrato di *Pc* essere espresso come: ${P_{i,c}}^2=(P_{i,c}-\frac{1}{3})^2-\frac{1}{9}+\frac{2}{3}P_{i,c}$; quindi sommando ottengo:

$$(1.3) \qquad {P_{i,a}}^2+{P_{i,b}}^2+{P_{i,c}}^2=(P_{i,a}-P_{i,b})^2+(P_{i,c}-\frac{1}{3})^2+2P_{i,a}P_{i,b}-\frac{1}{9}+\frac{2}{3}P_{i,c}$$

La somma delle frequenze relative è uguale a 1, pertanto vale anche per il suo quadrato, ovvero: $(P_{i,a}+P_{i,b}+P_{i,c})^2=1$; quindi $P_{i,a}^2+P_{i,b}^2+P_{i,c}^2+2P_{i,a}P_{i,b}+2P_{i,c}P_{1,a}+2P_{i,c}P_{i,b}=1$, sempre ricordando che la somma delle frequenze relative è pari all'unità ottengo:

$$(1.4) \qquad {P_{i,a}}^2+{P_{i,b}}^2+{P_{i,c}}^2=1-2P_{i,a}P_{i,b}-2P_{i,c}(1-P_{i,c})$$

Dalla (1.2) e dalla (1.4) ottengo: $\frac{d}{3N}+\frac{1}{3}<1-2P_{i,a}P_{i,b}-2P_{i,c}(1-P_{i,c})$ dalla quale posso estrarre il termine ottenendo: $2P_{i,a}P_{i,b}$

$$(1.5) \qquad 2P_{i,a}P_{i,b}<\frac{2}{3}-\frac{d}{3\,N}-2P_{i,c}(1-P_{i,c})$$

Dalla (1.2) e dalla (1.3) ottengo:

$$(1.6) \qquad \frac{d}{3N}+\frac{1}{3}<(P_{i,a}-P_{i,b})^2+\left(P_{i,c}-\frac{1}{3}\right)^2+2P_{i,a}P_{i,b}-\frac{1}{9}+\frac{2}{3}p_{i,c}$$

Rimpiazzando $2P_{i,a}P_{i,b}$ nella (1.6) con la (1.5), allora

$$\frac{d}{3\,N}+\frac{1}{3}<(P_{i,a}-P_{i,b})^2+(P_{i,c}-\frac{1}{3})^2+\frac{2}{3}-\frac{d}{3\,N}-2P_{i,c}(1-P_{i,c})-\frac{1}{9}+\frac{2}{3}P_{i,c}$$

Sommando e completando i quadrati ottengo:

$$(1.7) \qquad (P_{i,a}-P_{i,b})^2+3(P_{i,c}-\frac{1}{3})^2>\frac{2d}{3\,N}$$

In termini delle coordinate del grafico Triangolare, allora la (1.7) può essere riscritta segue:

$$(1.8) \qquad \frac{3\,N}{2d}X_i^2+\frac{9\,\mathrm{N}}{2d}(Y_i-\frac{1}{3})^2 > \ 1$$

la (1.8) eguagliata all'unità è l'equazione di una ellisse con semiassi

$$(1.9) \qquad Ax=\sqrt{\frac{2d}{3\,N}}\ ;\ Ay=\sqrt{\frac{2d}{9\,\mathrm{N}}}$$

Tutti i punti che si collocano esternamente all'ellisse presentano distribuzioni delle risposte diverse in modo statisticamente significative dalla distribuzione uniforme.

Appendice dim.2: cerchio di omogeneità

$$(2.1) \qquad N\left(\frac{(P_{i,a}-\frac{1}{3})^2}{1/3}+\frac{(P_{i,b}-\frac{1}{3})^2}{1/3}+\frac{(P_{i,c}-\frac{1}{3})^2}{1/3}\right)>d$$

Ricordando che $P_{i,a}+P_{i,b}+P_{i,c}=1$, la (2.1) può essere semplificata e riformulata come segue:

$$(2.2) \qquad P_{i,a}^2+P_{i,c}^2+\left(1-P_{i,a}-P_{i,c}\right)^2>\frac{d}{3N}+\frac{1}{3}$$

Sviluppando il quadrato e semplificando si perviene alla seguente disequazione:

$$(2.3) \qquad P_{i,a}^2+P_{i,c}^2-P_{i,a}-P_{i,c}+P_{i,a}P_{i,c}+\frac{1}{3}>\frac{d}{6N}$$

Dobbiamo individuare nel diagramma di coordinate $[X=P_{i,a}+P_{i,c}/2;\ Y=P_{i,c}\sqrt{3}/2]$ l'insieme dei punti che soddisfano la diseguaglianza (2.3), tale insieme sarà centrato sul punto $[1/2;\ \sqrt{3}/6]$ che esprime la distribuzione uniforme tra le categorie.

Osservando che:

$$(2.4) \qquad \left(P_{i,a}+\frac{P_{i,c}}{2}-\frac{1}{2}\right)^2=P_{i,a}^2+\frac{P_{i,c}^2}{4}+\frac{1}{4}+P_{i,a}P_{i,c}-P_{i,a}-\frac{P_{i,c}}{2}$$

$$(2.5) \qquad \left(P_{i,c}\frac{\sqrt{3}}{2}-\frac{\sqrt{3}}{6}\right)^2=P_{i,c}^2\frac{3}{4}-\frac{P_{i,c}}{2}+\frac{1}{12}$$

Sommando membro a membro la (2.4) e la (2.5) otteniamo

$$\left(P_{i,a}+\frac{P_{i,c}}{2}-\frac{1}{2}\right)^2+\left(P_{i,c}\frac{\sqrt{3}}{2}-\frac{\sqrt{3}}{6}\right)^2=P_{i,a}^2+P_{i,c}^2-P_{i,a}-P_{i,c}+P_{i,a}P_{i,c}+\frac{1}{3}$$

Quindi

$$\left(P_{i,a}+\frac{P_{i,c}}{2}-\frac{1}{2}\right)^2+\left(P_{i,c}\frac{\sqrt{3}}{2}-\frac{\sqrt{3}}{6}\right)^2>\frac{d}{6N} \quad (2.6)$$

Ricordando le coordinate del diagramma Trilineare $[X = P_{i,a} + P_{i,c}/2;\ Y = P_{i,c}\sqrt{3}/2]$ la (2.5) può essere riscritta

$$\left(X-\frac{1}{2}\right)^2+\left(Y-\frac{\sqrt{3}}{6}\right)^2>\frac{d}{6N} \quad (2.7)$$

La (2.7) è l'equazione di un cerchio di raggio:

$$r=\sqrt{\frac{d}{6N}} \quad (2.8)$$

Tutti i punti che si collocano esternamente al cerchio presentano distribuzioni interne delle risposte diverse in modo statisticamente significative dalla distribuzione uniforme.

Appendice dim.3: area di omogeneità parabolica.

Il test asintotico di Wald ha la seguente forma,

$$\chi^2_{(df=1)}\approx\frac{X^2}{Var(\mathrm{X})}>d \quad (3.1)$$

Dove *d* è il valore soglia del test del Chi Quadro con un grado di libertà al livello α%; mentre la varianza di $X_i = P_{i,a} - P_{i,b}$ può essere calcolata come secondo l'usuale formula: $Var(P_{i,a}-P_{i,b}) = Var(P_{i,a}) + Var(P_{i,b}) - 2Cov(P_{i,b}P_{i,a})$ dove $Cov(P_{i,a};P_{i,b}) = -\frac{P_{i,a}P_{i,b}}{N}$ [Zelterman2004]; pertanto la varianza della differenza assume la forma:

$$Var\left(P_{i,a}-P_{i,b}\right)=\frac{P_{i,a}(1-P_{i,a})}{N}+\frac{P_{i,b}(1-P_{i,b})}{N}+\frac{2}{N}P_{i,a}P_{i,b} \quad (3.2)$$

Riarrangiando la (3.2) si perviene alla

$$Var(P_{i,a}-P_{i,b})=\frac{(P_{i,a}-P_{i,b})-(P_{i,a}-P_{i,b})^2}{N} \quad (3.3)$$

Ricordando che $P_{i,a}-P_{i,b}=1-P_{i,c}$ e che $\mathrm{Y}_i = P_{i,c}$

$$Var\left(P_{i,a}-P_{i,b}\right)=\frac{1}{N}((1-Y_i)-X_i^2) \quad (3.4)$$

Quindi applicando l'equazione (3.4) alla (3.1) si perviene:

$$(3.5) \qquad \frac{NX_i^2}{[(1-Y_i)-X_i^2]} > d$$

Riarrangiando la (3.5) otteniamo

$$(3.6) \qquad Y_i > 1 - X_i^2\left(\frac{N+d}{d}\right)$$

La disequazione (3.6) delinea l'area esterna alla parabola che individua le posizioni per le quali le differenze tra la proporzioni Pa e Pb risultano statisticamente significative. Inoltre, risulta facile mostrare che la parabola (3.6) ha il suo massimo nel punto di coordinate [0;1] e interseca l'asse delle X nei punti di coordinate: $\left[-\sqrt{\frac{d}{N+d}};\, 0\right]$ e $\left[\sqrt{\frac{d}{N+d}};\, 0\right]$. Inoltre, il grafico triangolare è delimitato a sinistra (X<0) dal lato di equazione Y=-X+1 e a destra (X>0) dal lato di equazione Y=X+1, quindi è facile mostrare che la parabola interseca i lati del triangolo nei punti di coordinate: $\left[-\frac{d}{N+d};\, \frac{N}{N+d}\right]$ e $\left[\frac{d}{N+d};\, \frac{N}{N+d}\right]$

Appendice Codice R (Vers. 4.5.0) Grafico Triangolare

```
library(ggplot2)
file_path <- "C:/directory/ tabella2_dati.txt";
tabella2_dati <- read.delim(file_path, header = TRUE)
# Forza le colonne a numeriche (le prime tre)
tabella2_dati$SI     <- as.numeric(tabella2_dati$SI)
tabella2_dati$NO     <- as.numeric(tabella2_dati$NO)
tabella2_dati$NONSO  <- as.numeric(tabella2_dati$NONSO)
# Calcolo proporzioni e coordinate
N <- rowSums(tabella2_dati[, c("SI", "NO", "NONSO")])
Pa <- tabella2_dati$SI / N ;
Pb <- tabella2_dati$NO / N
Pc <- tabella2_dati$NONSO / N
X <- Pa – Pb; Y <- Pc
# Aggiunta colonne
tabella2_dati$Pa   <- Pa
tabella2_dati$Pb   <- Pb
tabella2_dati$Pc   <- Pc
tabella2_dati$X    <- X; tabella2_dati$Y    <- Y
tabella2_dati$Item <- paste0("Item", seq_len(nrow(tabella2_dati)))
# Parametri statistici (N item = num camp;  alpha = 0,05/12-confronti)
N_item <- 150; alpha <- 0.00417 #N numerosità campionaria; alpha =
0,05/"numero item"
d_ellisse  <- qchisq(1 - alpha, df = 2)
d_parabola <- qchisq(1 - alpha, df = 1)
# Calcolo dei semiassi dell'ellisse
a_ell <- sqrt((2/3) * d_ellisse / N_item)  # semiasse maggiore orizzontale (X)
b_ell <- sqrt((2/9) * d_ellisse / N_item)    # semiasse minore verticale (Y)
# Genera punti dell'ellisse parametrica
t <- seq(0, 2 * pi, length.out = 500)
ellisse <- data.frame(X = a_ell * cos(t),  Y = 1/3 + b_ell * sin(t))  # centrata
su Y = 1/3
# Parabola – disegna la parabola
a <- (d_parabola+N_item) / d_parabola
linf<- -sqrt(d_parabola/(N_item+d_parabola));
lsup<- sqrt(d_parabola/(N_item+d_parabola))
x_seq <- seq(linf, lsup, length.out = 1000)
parabola <- data.frame(  X = x_seq,  Y = 1-a *(x_seq^2))
ggplot() +
# Ellisse
geom_polygon(data = ellisse, aes(x = X, y = Y), fill = NA, color = "blue") +
```

```
# Parabola
geom_line(data = parabola, aes(x = X, y = Y), color = “red”) +
# Assi orizzontale e verticale
geom_hline(yintercept = 0, color = “gray30”, linewidth = 0.4) +
geom_vline(xintercept = 0, color = “gray30”, linewidth = 0.4) +
# Lati del triangolo
geom_segment(aes(x = -1, y = 0, xend = 0, yend = 1), color = “gray30”,
linewidth = 0.4) +
geom_segment(aes(x =  1, y = 0, xend = 0, yend = 1), color = “gray30”,
linewidth = 0.4) +
# Punti e etichette
geom_point(data = tabella2_dati, aes(x = X, y = Y), size = 2) +
geom_text(data = tabella2_dati, aes(x = X, y = Y, label = Item), nudge_y =
0.02, size = 3) +
# Titoli e tema
labs(title = “Grafico Triangolare con Ellisse e Parabola”, x = “a favore di (b)
<------ Pa - Pb ------> a favore di (a)”, y = “Pc”) +
scale_x_continuous(breaks = seq(-1, 1, by = 0.1), limits = c(-1, 1)) +
scale_y_continuous(breaks = seq(0, 1, by = 0.1), limits = c(0, 1)) +
theme_minimal() +
theme(axis.title.x = element_text(hjust = 0.5), plot.title = element_text(h-
just = 0.5) )
```