

Análisis de concordancia de atributos

Fecha de recepción: 21/07/08

Fecha de aceptación: 05/09/08

Federico Picado Alvarado¹

Palabras clave

Atributos, características de calidad, concordancia, binario, nominal, ordinal, Kappa, Kendall, Minitab.

Key words

Attributes, quality characteristics, binary, nominal, ordinal, Kappa, Kendall, Minitab.

Resumen

Se presenta la metodología del análisis de concordancia de atributos para la evaluación de las respuestas de los evaluadores, cuando se analizan características de calidad tipo atributo, donde la subjetividad puede afectar la efectividad del sistema de medición. El análisis de concordancia de atributos puede ser muy útil en sistemas de calidad donde se aplique métodos de inspección por atributos, tanto en organizaciones de manufactura como en servicios. Se desarrolla un ejemplo de aplicación con el uso del software estadístico Minitab 15.1.

Abstract

An attribute agreement analysis assesses the consistency of responses within appraisers and between appraisers. It also compares, if appropriate the responses with “master” or “reference” values (also called standard values). The analysis uses attribute ratings or classifications. The data can be binary, nominal or ordinal. An attribute agreement analysis is useful in quality systems where inspection procedures based on attributes are applied in manufacturing and service organizations. In addition, an application example is developed using Minitab 15.1

Introducción

Un análisis de concordancia de atributos evalúa la uniformidad de las respuestas dentro de un grupo de evaluadores y entre ellos y, en caso de ser apropiado, compara las respuestas con los valores “patrón” o “referenciales” (también denominados “valores estándares”). El análisis utiliza las calificaciones o clasificaciones de los atributos. Las mediciones son evaluaciones subjetivas hechas por personas y no por

1. Profesor de la Escuela de Ingeniería en Producción Industrial del Instituto Tecnológico de Costa Rica. Teléfono: 2550-2275. Correo electrónico: fpicado@itcr.ac.cr.



mediciones físicas directas. Algunos ejemplos pueden ser:

- Evaluaciones del rendimiento de un automóvil.
- Clasificación de la calidad de una tela como “buena” o “mala”.
- Clasificación del sabor, aroma y color de un vino en una escala de 1 a 10.

En estas situaciones, las características de calidad son difíciles de definir y evaluar. Para obtener clasificaciones significativas, más de un evaluador debe coincidir en la medición de la respuesta. Si los evaluadores están de acuerdo, existe la posibilidad de que las calificaciones sean precisas. Si los evaluadores no concuerdan, la utilidad de la clasificación es limitada.

Los datos tipo atributo pueden ser binarios, nominales u ordinales, a saber:

Tipo	Datos	Ejemplos
Binarios	Uno de dos valores	<ul style="list-style-type: none"> • La respuesta puede ser correcta o incorrecta. • La parte puede ser rechazada o aceptada.
Nominales	Variables categóricas no ordenadas	<ul style="list-style-type: none"> • La tela tiene defectos como rasgaduras, fallas de costura o problemas de estampado. • El sabor del vino es seco, dulce o afrutado.
Ordinales	Variables discretas ordenadas	<ul style="list-style-type: none"> • Las medallas olímpicas son oro, plata y bronce. • La calidad de un vino se califica en una escala de 1 a 5. • Respuestas que tengan tres o más niveles con un orden natural, tales como: en desacuerdo fuertemente, en desacuerdo, neutral, de acuerdo, fuertemente de acuerdo.

El análisis de concordancia de atributos se puede utilizar para evaluar la uniformidad o exactitud de las respuestas sobre atributos cuando:

- Dos o más evaluadores o instrumentos califiquen los mismos elementos.

- El mismo evaluador o instrumento califique los mismos elementos más de una vez.
- Un evaluador califique los elementos una vez (sólo con fines de exactitud).

El análisis de concordancia de atributos responde a preguntas como:

- ¿Son uniformes las respuestas de un evaluador?
- ¿Califican los diferentes evaluadores los mismos elementos de manera uniforme (es decir, existe concordancia entre los evaluadores)?
- ¿Son exactas las respuestas de los evaluadores cuando se comparan con un estándar conocido o una respuesta correcta?
- ¿Es efectivo el sistema de medición de atributos?
- ¿Concuerdan los operarios en las calificaciones pasa/no pasa respecto a la muestra de un producto terminado?
- ¿Son correctas las calificaciones de los operarios cuando se comparan con el estándar suministrado por el coordinador de calidad?
- ¿Es efectivo un sistema de inspección visual para detectar cuándo falta algún componente clave en un ensamblaje?

El análisis adecuado de datos tipo atributo es muy común y necesario en varios sistemas de calidad en organizaciones de manufactura y servicios. Sin embargo, a veces se ha dejado de lado el uso de herramientas estadísticas diseñadas para lograr una mejor interpretación de los resultados y así poder apoyar mejor la toma de decisiones.

Análisis Estadístico Recomendado

Estadísticas Kappa de Fleiss

El valor Kappa representa el grado de concordancia absoluta entre las

calificaciones, es decir, Kappa trata todas las clasificaciones incorrectas de igual forma, sin tomar en cuenta su magnitud. Se utiliza las estadísticas Kappa cuando sus clasificaciones son binarias, nominales u ordinales; sin embargo, debido a que Kappa no explica la magnitud de la diferencia observada, es probable que también deba considerarse otras estadísticas en su análisis.

Interpretación de las estadísticas Kappa

La hipótesis nula es que Kappa es igual a 0, lo cual indicaría que el nivel de concordancia es el mismo que el esperado en virtud de las probabilidades. Los valores de Kappa tienen un rango de -1 a +1. Mientras más fuerte sea la concordancia absoluta entre las calificaciones, más alto será el valor de Kappa:

- Si $Kappa = 1$, la concordancia es perfecta.
- Si $Kappa = 0$, la concordancia es igual a la esperada en virtud de las probabilidades.
- Si $Kappa < 0$, la concordancia es más débil que la esperada en virtud de las probabilidades (Esto ocurre muy pocas veces).

Las directrices de la AIAG [1] sugieren que “valores de Kappa mayores que 0.75 indican una concordancia de buena a excelente (con un Kappa máximo = 1); los valores menores que 0.40 indican poca concordancia”.

Coefficiente de Concordancia de Kendall

Cuando se tiene datos de tipo ordinal tales como la clasificación de la severidad de los defectos en una escala de 1 a 5, el coeficiente de Kendall, el cual toma en consideración el orden, es usualmente un instrumento estadístico más apropiado para evaluar la asociación que únicamente el Kappa.

El coeficiente de Kendall indica el grado de asociación de las evaluaciones ordinales

hechas por evaluadores múltiples cuando se evalúa las mismas muestras. Los valores del coeficiente de Kendall tienen un rango de 0 a +1. Entre mayor sea el valor Kendall, más fuerte será la asociación. Generalmente, coeficientes Kendall de 0.9 ó mayores son considerados muy buenos. Un coeficiente Kendall alto o significativo implica que los evaluadores están aplicando esencialmente el mismo estándar cuando evalúan las muestras.

Análisis de concordancia de atributos con datos binarios

Ejemplo de aplicación: Sustratos de convertidor catalítico

Se utilizó las funcionalidades del software Minitab 15.1 para el análisis del ejemplo de aplicación.

Problema

Una compañía de cerámica prepara sustratos de convertidor catalítico para un proveedor automotor. El sustrato cerámico es una estructura en forma de panal sobre la que se apoya el material catalizador que reduce las emisiones de escape.

En las fases finales de producción, un evaluador verifica que no existan fisuras en cada sustrato. Una parte con fisuras se considera defectuosa y debe desecharse.

Los evaluadores colocan cada parte bajo una luz para inspeccionar si existen fisuras. El equipo de fabricación desea verificar que sus evaluadores nuevos califiquen las partes en forma efectiva. El Departamento de Calidad ha evaluado las muestras y suministrado los valores estándares para cada parte.

Recolección de datos

Cuatro evaluadores inspeccionan diez partes en orden aleatorio y registran las respuestas, bien sea como “Pasa” o “No pasa”. Luego de que las partes se ordenan nuevamente, cada evaluador las inspecciona una vez más.

... El análisis de concordancia de atributos puede ser útil en sistemas de calidad donde se emplee métodos de aplicación por atributos, tanto en organizaciones de manufactura como en servicios.

Variable	Descripción
Respuesta	Calificación del evaluador del sustrato (Pasa/No Pasa)
Parte	Sustrato evaluado
Evaluador	Evaluador que calificó el sustrato
Orden Corrida	Orden original de las corridas del estudio estándar
Estándar	Calificación correcta para cada sustrato

La compañía de cerámica desea evaluar:

- La uniformidad de las respuestas dentro de cada evaluador.
- La uniformidad entre diferentes evaluadores al calificar las mismas muestras.
- La exactitud de las calificaciones de los evaluadores, de acuerdo con el estándar.

Para determinar si los evaluadores califican de forma correcta, debe suministrarse una columna (Estándar) con el estándar conocido para cada muestra.

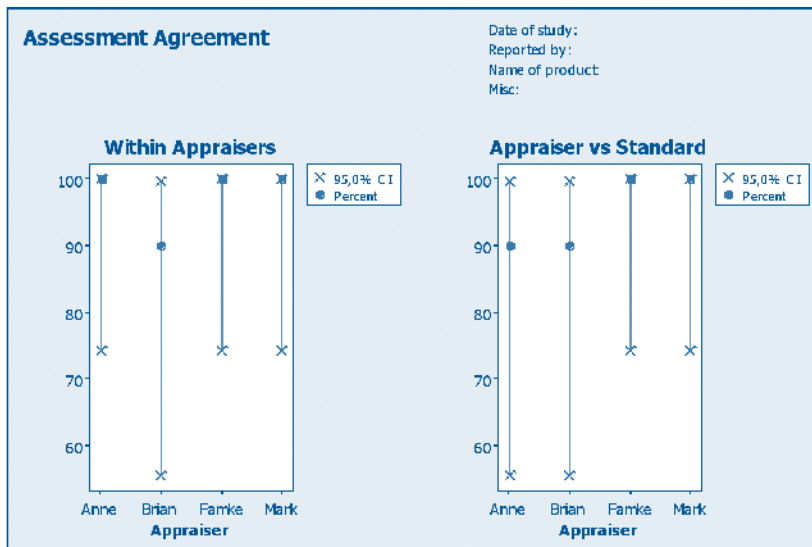


Figura 1. Uniformidad de las respuestas de cada evaluador.

Interpretación de sus resultados

Para los evaluadores

La Figura.1 muestra la uniformidad de las respuestas de cada evaluador. Anne, Famke y Mark coincidieron las dos veces cuando calificaron las diez partes, para una tasa de coincidencia del 100%. Brian tuvo un problema de concordancia al calificar una de las diez partes, para una tasa de coincidencia del 90%.

Evaluador versus estándar

Esta gráfica muestra cómo los resultados de cada evaluador se comparan con el estándar. Famke y Mark calificaron con un 100% de exactitud, mientras que Anne y Brian calificaron nueve de las diez partes de manera correcta ambas veces, para una tasa de exactitud del 90%.

Tamaño de los intervalos de confianza

Se estima que la habilidad de Brian para coincidir consigo mismo en dos evaluaciones de la misma parte es de un 90%, con un intervalo de confianza del 95%, que va de 55,5% a 99,75%. Este intervalo es muy grande debido a que el tamaño de la muestra es pequeño (diez partes). Un estudio de concordancia de atributos con 30 ó 50 partes produciría resultados más precisos e informativos.

Interpretación de sus resultados

Para el evaluador

Si cada evaluador inspecciona la misma parte dos o más veces, usted puede evaluar la uniformidad de las respuestas para cada evaluador.

De los cuatro evaluadores, sólo Brian suministró una respuesta mixta para una de las diez partes. En la hoja de trabajo (datos del problema no mostrados) se observó que Brian registró “Pasa” para la Parte #1 cuando la evaluó por primera vez, pero “No pasa”, la segunda vez. Los otros tres evaluadores, Anne, Famke y Mark

registraron respuestas uniformes para todas las partes en ambos ensayos. Los valores Kappa indican una fuerte concordancia “dentro” de cada operador. Sólo Brian, con una respuesta mixta, posee un valor Kappa menor que 1 (concordancia perfecta).

Assessment Agreement				
Appraiser	# Inspected	# Matched	Percent	95 % CI
Anne	10	10	100,00	(74,11. 100,00)
Brian	10	9	90,00	(55,50. 99,75)
Famke	10	10	100,00	(74,11. 100,00)
Mark	10	10	100,00	(74,11. 100,00)
# Matched: Appraiser agrees with him/herself across trials.				

Fleiss' Kappa Statistics						
Appraiser	Response	Kappa	SE Kappa	Z P(vs > 0)		
Anne	No pasa	1,00000	0,316228	3,16228	0,0008	
	Pasa	1,00000	0,316228	3,16228		0,0008
Brian	No pasa	0,78022	0,316228	2,46727		0,0068
	Pasa	0,78022		0,316228	2,46727	0,0068
Famke	No pasa	1,00000	0,316228	3,16228	0,0008	
	Pasa	1,00000	0,316228	3,16228	0,0008	
Mark	No pasa	1,00000	0,316228	3,16228	0,0008	
	Pasa	1,00000	0,316228	3,16228	0,0008	

Interpretación de sus resultados

Cada evaluador *versus* estándar: concordancia de evaluaciones

Debido a que se provee un valor estándar para cada parte, se puede comparar los

resultados de los evaluadores con los valores “reales”.

Anne y Brian calificaron nueve de las diez partes correctamente ambas veces; Famke y Mark calificaron las diez partes correctamente de manera uniforme.

Cada evaluador *versus* estándar: discrepancia en la evaluación

La tabla denominada Discrepancia en la evaluación indica la discrepancia del operador con el estándar para cada operador.

- La columna # Pasa / No pasa contiene el número de partes que el operador pasó, pero que no debieron haber pasado (un falso positivo).
- La columna # No pasa / Pasa contiene el número de partes que el operador no pasó, pero que debieron haber pasado (un falso negativo).
- La columna # Mixed indica cuántas veces el operador no coincidió consigo mismo durante los ensayos.

En ambos ensayos, Anne pasó una parte de cuatro que no debió haber pasado (un nivel de significancia del 25%). Brian no coincidió consigo en una parte (una respuesta mixta para una de diez partes o nivel de significancia del 10%).

Se debe consultar la hoja de trabajo para identificar los valores originales registrados que causaron estos errores. Por ejemplo, Brian suministró una respuesta mixta para el sustrato #1 y Anne indicó un falso positivo para el sustrato #2. Los valores Kappa indican que sólo Anne (0,78022) y Brian (0,89011) cometieron errores en comparación con el estándar.

Assessment Agreement				
Appraiser	# Inspected	# Matched	Percent	95 % CI
Anne	10	9	90,00	(55,50. 99,75)
Brian	10	9	90,00	(55,50. 99,75)
Famke	10	10	100,00	(74,11. 100,00)
Mark	10	10	100,00	(74,11. 100,00)

Matched: Appraiser's assessment across trials agrees with the known standard.

Assessment Disagreement						
	# Pasa		# No pasa			
Appraiser	No pasa	Percent	Pasa	Percent	# Mixed Percent	
Anne	1	25,00	0	0,00	0	0,00
Brian	0	0,00	0	0,00	1	10,00
Famke	0	0,00	0	0,00	0	0,00
Mark	0	0,00	0	0,00	0	0,00

Pasa / No pasa: Assessments across trials = Pasa / standard = No pasa.
 # No pasa / Pasa: Assessments across trials = No pasa / standard = Pasa.
 # Mixed: Assessments across trials are not identical.

Fleiss' Kappa Statistics						
Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)	
Anne	No pasa	0,78022	0,223607	3,48925	0,0002	
	Pasa	0,78022	0,223607	3,48925	0,0002	
Brian	No pasa	0,89011	0,223607	3,98069	0,0000	
	Pasa	0,89011	0,223607	3,98069	0,0000	
Famke	No pasa	1,00000	0,223607	4,47214	0,0000	
	Pasa	1,00000	0,223607	4,47214	0,0000	
Mark	No pasa	1,00000	0,223607	4,47214	0,0000	
	Pasa	1,00000	0,223607	4,47214		0,0000

Entre los evaluadores

Los cuatro evaluadores coincidieron en sus calificaciones para ocho de diez partes inspeccionadas. Sin embargo, esta tabla sólo indica la uniformidad entre los operadores, no la exactitud; no evalúa

cómo sus respuestas se comparan con los estándares.

El valor Kappa (0.853183) para ambas respuestas indica una concordancia adecuada entre los evaluadores.

Assessment Agreement			
# Inspected	# Matched	Percent	95 % CI
10	8	80,00	(44,39. 97,48)
# Matched: All appraisers' assessments agree with each other.			

Fleiss' Kappa Statistics				
Response	Kappa	SE Kappa	Z	P(vs > 0)
No pasa	0,853183	0,0597614	14,2765	0,0000
Pasa	0,853183	0,0597614	14,2765	0,0000

Todos los evaluadores versus estándar

Estos resultados muestran que los cuatro evaluadores coincidieron de forma correcta con las calificaciones estándares de ocho de las diez partes.

El valor Kappa (0.917582) para ambas respuestas indica una fuerte concordancia con los valores estándares.

Assessment Agreement			
# Inspected	# Matched	Percent	95 % CI
10	8	80,00	(44,39. 97,48)
# Matched: All appraisers' assessments agree with the know standard.			

Fleiss' Kappa Statistics				
Response	Kappa	SE Kappa	Z	P(vs > 0)
No pasa	0,917582	0,111803	8,20711	0,0000
Pasa	0,917582	0,111803	8,20711	0,0000

Consideraciones finales

Resumen y conclusiones

- En ambos ensayos, dos de los cuatro evaluadores calificaron las diez partes correctamente.
- Brian calificó una parte de manera discrepante, es decir, sus mediciones en ambos ensayos no coincidieron.
- Anne indicó un falso positivo para una parte. Se debería investigar con más detalle por qué ocurrieron estos errores.

Bibliografía

1. AIAG *Measurement System Analysis (MSA) Reference Manual*, 3rd edition (2002). Automotive Industry Action Group.
2. AIAG *Statistical Process Control (SPC) Reference Manual* (1995). Automotive Industry Action Group.
3. Y. Chou, A.M. Polansky y R.L. Mason (1998). "Transforming Nonnormal Data to Normality in Statistical Process Control", *Journal of Quality Technology*, 30.
4. W. Edwards Deming (1990). Foreword to *Statistical Method from the Viewpoint of Quality Control*, by Walter A. Shewhart. Dover Publications, Inc.
5. D.C. Montgomery (2000). *An Introduction to Statistical Quality Control*, 4th edition. John Wiley & Sons, Inc.
6. S.E. Somerville y D.C. Montgomery (1996). "Process Capability Indices and Non-Normal Distributions," *Quality Engineering*, Vol. 9, No. 2.
7. *Statistical Process Control Reference Manual*, 1st Edition (1995), Chrysler, Ford, and General Motors Corporation, p.79– 80.