variable categórica – categorical variable

Authored by

memjavad

November 12, 2025

Variable Categórica

Primary Disciplinary Field(s): Estadística, Ciencia de Datos, Investigación Cuantitativa

1. Definición Fundamental y Escalas de Medición

Una **variable categórica**, también conocida como variable cualitativa o nominal, constituye un pilar fundamental en la estadística descriptiva e inferencial. Se define como aquella variable cuyos valores observados o registrados representan categorías o grupos, en lugar de cantidades medibles en una escala numérica continua. La esencia de la variable categórica radica en su capacidad para clasificar una unidad de análisis (como un individuo, evento u objeto) dentro de uno y solo uno de un conjunto finito de clases mutuamente excluyentes y exhaustivas. A diferencia de las variables cuantitativas, las operaciones aritméticas como la suma o la resta carecen de sentido intrínseco para los valores de una variable categórica, ya que sus "valores" son etiquetas o nombres que denotan pertenencia a un grupo.

La comprensión de la variable categórica está intrínsecamente ligada al concepto de escalas de medición, propuesto originalmente por el psicólogo Stanley Smith Stevens en 1946. Las variables categóricas ocupan los niveles más bajos de esta jerarquía: la escala nominal y la escala ordinal. La distinción entre estas escalas es crucial, ya que determina el tipo de análisis estadístico apropiado. Mientras que la escala nominal solo permite la identificación y la clasificación (por ejemplo, el color de ojos o la nacionalidad), la escala ordinal introduce una relación de orden o jerarquía entre las categorías (por ejemplo, niveles de satisfacción o grados académicos), aunque la distancia entre dichas categorías no es cuantificable ni uniforme.

El manejo adecuado de las variables categóricas es indispensable en la investigación social, médica y de mercado. Su correcta identificación previene el uso de métodos estadísticos inapropiados que podrían llevar a conclusiones erróneas. Por ejemplo, intentar calcular la media de un conjunto de códigos numéricos asignados arbitrariamente a categorías nominales (como 1=Hombre, 2=Mujer) resulta en un valor sin significado contextual. Por lo tanto, el rigor estadístico exige que el investigador seleccione técnicas basadas en frecuencias, proporciones, o pruebas no paramétricas cuando se enfrenta a este tipo de datos.

2. Tipos de Variables Categóricas: Nominal

La **variable nominal** representa la forma más simple de variable categórica. Su característica definitoria es que las categorías que la componen no poseen ningún orden intrínseco, jerarquía o secuencia lógica. La única operación válida que se puede realizar con estos datos es determinar si dos observaciones son iguales o diferentes. Los números o símbolos que se asignan a estas categorías sirven únicamente como identificadores o etiquetas. Si se codifica la variable "Estado Civil" como 1=Soltero, 2=Casado, 3=Divorciado, el número 3 no implica que "Divorciado" sea

"más" o "mejor" que "Soltero"; es simplemente un marcador distintivo.

Ejemplos comunes de variables nominales incluyen el sexo (masculino, femenino), el grupo sanguíneo (A, B, AB, O), la afiliación política o la región geográfica. En el contexto de la ciencia de datos, las variables nominales suelen requerir transformaciones específicas, como la **codificación *one-hot***, antes de poder ser utilizadas en algoritmos de aprendizaje automático. Esta codificación convierte la variable nominal en un conjunto de variables binarias (o *dummy variables*), donde cada nueva variable representa una categoría y toma el valor 1 si la observación pertenece a esa categoría y 0 en caso contrario. Este proceso es vital para evitar que los algoritmos interpreten erróneamente las etiquetas numéricas como valores cuantitativos.

Dentro del espectro nominal, existe un subtipo particularmente importante: la **variable dicotómica** o binaria. Esta variable es aquella que solo puede tomar dos posibles valores o categorías, como "Sí/No", "Verdadero/Falso" o "Éxito/Fracaso". Aunque es nominal, la variable dicotómica posee propiedades matemáticas únicas que facilitan su uso en modelos de regresión, como la regresión logística, donde el resultado de interés es precisamente la probabilidad de pertenecer a una de las dos categorías. La simplicidad de la estructura dicotómica permite una interpretación directa de los coeficientes del modelo, representando el cambio en el logaritmo de las probabilidades de ocurrencia del evento.

3. Tipos de Variables Categóricas: Ordinal

La **variable ordinal** se distingue de la nominal por la presencia de un orden significativo entre sus categorías. Si bien las categorías son distintas y mutuamente excluyentes, existe una progresión lógica o una jerarquía que permite al investigador clasificar las observaciones de menor a mayor o viceversa. Sin embargo, y esta es la limitación crucial, la magnitud de la diferencia o la distancia entre las categorías adyacentes no puede asumirse como igual o constante. Por ejemplo, en una escala de Likert de 5 puntos (1=Totalmente en desacuerdo, 5=Totalmente de acuerdo), la diferencia percibida entre el punto 1 y 2 no es necesariamente la misma que la diferencia entre el punto 4 y 5.

La naturaleza ordenada de estas variables permite el cálculo de estadísticas basadas en el rango, como la mediana y los percentiles, que son inapropiados para variables nominales. En el análisis inferencial, las variables ordinales a menudo requieren el uso de pruebas no paramétricas, como la prueba de Kruskal-Wallis o la prueba de Wilcoxon, que se basan en la ordenación de los datos en lugar de asumir una distribución normal o la igualdad de varianzas. La aplicación incorrecta de pruebas paramétricas (diseñadas para datos de intervalo o razón) a datos ordinales puede distorsionar los resultados, ya que estas pruebas asumen que las distancias entre los puntos de la escala son equivalentes.

Ejemplos típicos de variables ordinales se encuentran frecuentemente en la medición de actitudes,

opiniones y clasificaciones socioeconómicas. Esto incluye los niveles de educación (primaria, secundaria, universitaria), la severidad de una enfermedad (leve, moderada, severa) o las calificaciones de calidad de un producto (malo, regular, bueno, excelente). Cuando se utilizan variables ordinales en modelos predictivos avanzados, a menudo se manejan de manera similar a las variables nominales mediante codificación, pero algunos métodos estadísticos avanzados intentan explotar la información de orden inherente, como los modelos de regresión logística ordinal, que son específicamente diseñados para este tipo de estructura de datos.

4. Representación y Codificación de Datos Categóricos

La transición de los datos categóricos de su formato cualitativo original a un formato que pueda ser procesado por algoritmos matemáticos constituye uno de los pasos más importantes en la preparación de datos. Si bien los humanos interpretamos fácilmente etiquetas como "Rojo" o "Azul", la mayoría de los modelos estadísticos y de aprendizaje automático requieren entradas numéricas. La técnica de codificación más fundamental es la **codificación de etiquetas** (*label encoding*), donde a cada categoría única se le asigna un número entero arbitrario (e.g., Rojo=1, Azul=2, Verde=3). Esta técnica es simple, pero peligrosa si se aplica a variables nominales, ya que introduce falsas relaciones ordinales que el modelo podría interpretar como jerarquías cuantitativas.

Para mitigar el riesgo de inferir un orden inexistente en las variables nominales, se utiliza predominantemente la **codificación *One-Hot*** (codificación por variables indicadoras o ficticias). Como se mencionó previamente, esta técnica expande la variable categórica original en 'k' nuevas variables binarias, donde 'k' es el número de categorías únicas. Si un registro pertenece a la categoría 'k', la variable binaria 'k' toma el valor 1, mientras que las otras 'k-1' variables toman el valor 0. Un aspecto crucial de la codificación One-Hot, especialmente en modelos de regresión lineal, es la necesidad de omitir una de las categorías binarias resultantes. Esta categoría omitida, denominada la **categoría de referencia** o línea base, es necesaria para evitar la multicolinealidad perfecta (el fenómeno de la trampa de la variable ficticia), asegurando que la matriz de diseño del modelo sea de rango completo y que los coeficientes sean interpretables en relación con esa base.

Otras técnicas de codificación más sofisticadas han surgido, particularmente en el ámbito del *machine learning*, para manejar variables categóricas con un número muy elevado de categorías (alta cardinalidad). Estas incluyen la **codificación de frecuencia** (donde la categoría se reemplaza por la frecuencia con la que aparece en el conjunto de datos) y la **codificación objetivo** (*target encoding*), donde la categoría se reemplaza por la media de la variable objetivo correspondiente a esa categoría. Estas técnicas buscan reducir la dimensionalidad y capturar la información predictiva de la variable categórica de manera más eficiente, aunque deben aplicarse con precaución para evitar el sobreajuste (*overfitting*), especialmente la codificación objetivo, que

puede introducir sesgos si no se utiliza con técnicas de validación cruzada adecuadas.

5. Análisis Estadístico Aplicado a Variables Categóricas

El análisis de variables categóricas se centra en la distribución de frecuencias y las relaciones de asociación, más que en la dispersión o la tendencia central cuantitativa. El método fundamental para visualizar y resumir estos datos es la tabla de frecuencias, que muestra el número y el porcentaje de observaciones que caen en cada categoría. Cuando se analizan dos variables categóricas simultáneamente, se utiliza la **tabla de contingencia** (o tabla cruzada), que organiza la frecuencia de las combinaciones de categorías de ambas variables, permitiendo la exploración de la dependencia o independencia entre ellas.

Para probar si existe una asociación estadísticamente significativa entre dos variables categóricas nominales, la prueba más común es la **prueba de ji-cuadrado** (Chi-cuadrado de Pearson). Esta prueba evalúa si las frecuencias observadas en la tabla de contingencia difieren significativamente de las frecuencias que se esperarían si las variables fueran completamente independientes. Si bien la prueba de ji-cuadrado informa sobre la existencia de una asociación, no cuantifica la fuerza de dicha relación. Para medir la fuerza de la asociación, se recurre a coeficientes como el coeficiente Phi (para tablas 2x2), la V de Cramer o el coeficiente de contingencia, que normalizan la medida de asociación.

En el contexto de las variables categóricas ordinales, el análisis debe aprovechar la información de orden. Las medidas de asociación basadas en rangos, como el **coeficiente Tau de Kendall** o el **coeficiente Gamma**, son más apropiadas que las medidas nominales, ya que consideran la concordancia y la discordancia de los pares de observaciones. Además, las técnicas de regresión no paramétrica o los modelos logísticos ordinales son esenciales para predecir la probabilidad de que una observación caiga en una categoría específica, respetando la estructura jerárquica de la variable dependiente. La selección precisa del método analítico es lo que garantiza que las inferencias extraídas de los datos categóricos sean válidas y significativas para el dominio de estudio.

6. Importancia en la Modelización Predictiva

Las variables categóricas desempeñan un papel crítico en la modelización predictiva, tanto en la estadística tradicional como en el aprendizaje automático. En la modelización lineal clásica, como la **regresión lineal múltiple**, las variables categóricas independientes se incorporan mediante la ya mencionada codificación de variables ficticias. Los coeficientes resultantes de estas variables ficticias miden el cambio promedio en la variable dependiente al pasar de la categoría de referencia a la categoría codificada, manteniendo todas las demás variables constantes. Esto permite la inclusión de factores cualitativos (como la marca del producto o el método de

tratamiento) en modelos que, de otro modo, solo manejarían variables cuantitativas.

En el campo del aprendizaje automático, las variables categóricas son esenciales en diversas tareas de clasificación. Por ejemplo, si el objetivo es predecir si un cliente abandonará un servicio (una variable dependiente binaria), las variables independientes categóricas como "tipo de contrato", "método de pago" o "ubicación geográfica" son cruciales para construir modelos de clasificación robustos como árboles de decisión, bosques aleatorios o máquinas de soporte vectorial. De hecho, los algoritmos basados en árboles son particularmente adecuados para manejar variables categóricas, ya que dividen el espacio de características basándose en las categorías, sin necesidad de la codificación One-Hot, simplificando el proceso de preprocesamiento.

Finalmente, la correcta interpretación de las variables categóricas en los modelos es vital para la interpretabilidad del modelo. En el análisis causal, el uso de variables categóricas permite identificar si un tratamiento, una exposición o una característica específica tiene un efecto diferenciado sobre el resultado. Esto es fundamental en campos como la epidemiología (efecto de un factor de riesgo categorizado), la economía (impacto de una política dividida por región) o la investigación de operaciones (eficiencia de diferentes líneas de producción). El valor de la variable categórica reside en su capacidad para segmentar la población y revelar patrones que podrían quedar ocultos si solo se utilizaran datos puramente cuantitativos.

7. Consideraciones Críticas y Limitaciones

A pesar de su ubicuidad, el manejo de variables categóricas presenta varias limitaciones y desafíos metodológicos que deben ser considerados por el analista. Uno de los problemas principales es la **pérdida de información** que ocurre cuando una variable intrínsecamente continua o de razón se reduce a una escala categórica. Por ejemplo, categorizar la edad (una variable de razón) en "joven", "adulto" y "mayor" facilita la interpretación, pero desecha la información detallada sobre las diferencias de edad dentro de cada grupo, reduciendo la potencia estadística del análisis. Esta categorización forzada debe realizarse solo cuando la teoría subyacente lo justifique o cuando la variable continua muestre severos problemas de distribución.

Otro desafío significativo es el manejo de la **alta cardinalidad**. Una variable categórica con cientos o miles de categorías únicas (por ejemplo, códigos postales, identificadores de producto) puede generar una explosión dimensional si se aplica la codificación One-Hot, creando un número excesivo de variables binarias dispersas. Esto no solo sobrecarga la memoria computacional, sino que también puede conducir al sobreajuste y a la inestabilidad de los modelos. La gestión de la cardinalidad alta requiere técnicas avanzadas como la agrupación de categorías raras, el uso de codificación objetivo o la aplicación de métodos de reducción de dimensionalidad específicos.

Finalmente, la interpretación de la asociación en las tablas de contingencia puede ser engañosa,

particularmente en presencia de **variables de confusión** o cuando el tamaño de la muestra es pequeño. La prueba de ji-cuadrado asume que el tamaño de las celdas esperadas no es demasiado pequeño; si esta suposición se viola, las conclusiones sobre la significancia estadística pueden ser erróneas. En tales casos, se debe recurrir a la prueba exacta de Fisher. Además, es crucial recordar que la asociación estadística, incluso cuando es fuerte, no implica causalidad; la variable categórica solo indica la segmentación y la relación de ocurrencia entre los grupos definidos.

8. Lecturas Adicionales

Variable estadística - Wikipedia
Nivel de medición (Escalas de Stevens) - Wikipedia
Variable ficticia (Dummy Variable) - Wikipedia
Categorical Variables - StatSoft Electronic Textbook