ANÁLISIS DE DATOS: DESCUBRIENDO LA DISITRIBUCIÓN IDEAL CON MÉTRICAS AVANZADAS

Herramientas de Ciencia de Datos II

Ashley Arrieta PAdilla
Universidad de Costa Rica

Prof: Luis Alberto Potoy

1 Introducción

El trabajo actuarial es una disciplina clave en el sector de seguros y finanzas, donde se requiere una precisa evaluación y gestión de riesgos. Los actuarios utilizan técnicas estadísticas y matemáticas para modelar y prever eventos futuros, estableciendo primas y reservas adecuadas para diversas pólizas de seguro. Dos elementos esenciales en este proceso son la correcta identificación de las distribuciones de datos y la imputación adecuada de datos faltantes.Además, se aborda la relevancia de comprender la distribución de los datos y la importancia de una correcta imputación de datos faltantes.

2 Métodos usados

En el análisis actuarial, seleccionar la distribución correcta para un conjunto de datos es crucial para la precisión de los modelos predictivos y las decisiones basadas en datos. Para evaluar la adecuación de diferentes distribuciones, se utilizan varios criterios estadísticos y pruebas. En esta sección, se explican en detalle tres métodos clave: el Criterio de Información de Akaike (AIC), el Criterio Bayesiano de Información (BIC), y la prueba de Kolmogorov-Smirnov.

Criterio de Información de Akaike (AIC)

El Criterio de Información de Akaike (AIC) es una medida utilizada para comparar la calidad de diferentes modelos de distribución. Desarrollado por Hirotsugu Akaike, el AIC no solo evalúa la bondad de ajuste de un modelo, sino que también penaliza la complejidad del modelo para evitar el sobreajuste. Un modelo con un AIC más bajo se considera mejor ajustado a los datos.

Fórmula del AIC

$$\text{AIC} = 2k - 2\ln(L)$$

Donde:

- k es el número de parámetros del modelo.
- L es la verosimilitud del modelo.

Criterio Bayesiano de Información (BIC)

El Criterio Bayesiano de Información (BIC) es similar al AIC pero introduce una penalización más fuerte para la complejidad del modelo. Es especialmente útil cuando se trabaja con grandes conjuntos de datos.

Fórmula del BIC

$$\text{BIC} = k\ln(n) - 2\ln(L)$$

Donde:

- k es el número de parámetros del modelo.

- n es el tamaño de la muestra.
- L es la verosimilitud del modelo.

El término $k \ln(n)$ penaliza la complejidad del modelo más fuertemente en comparación con el AIC, especialmente a medida que aumenta el tamaño de la muestra n.

Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov (KS) es una prueba no paramétrica utilizada para comparar una muestra con una distribución de referencia (prueba de una muestra) o para comparar dos muestras (prueba de dos muestras). Es especialmente útil para evaluar la bondad de ajuste de una distribución.

Procedimiento de la Prueba KS

1. **Calcular la Función de Distribución Empírica (FDE)**: La FDE se construye a partir de los datos observados.
2. **Calcular la Función de Distribución Teórica (FDT)**: La FDT se basa en la distribución de referencia.
3. **Calcular la estadística D**: La estadística D es la mayor diferencia absoluta entre la FDE y la FDT.

$$D = \sup_x |F_n(x) - F(x)|$$

Donde:

- $F_n(x)$ es la FDE.
- $F(x)$ es la FDT.

3 En Python

La clase `AjusteDistribuciones` está diseñada para realizar el ajuste y la comparación de distribuciones estadísticas a un conjunto de datos utilizando las herramientas proporcionadas por la biblioteca `scipy.stats`. Esta clase es especialmente útil en el campo actuarial, donde la identificación precisa de la distribución de los datos es crucial para el análisis y la toma de decisiones.

El propósito principal de la clase `AjusteDistribuciones` es automatizar el proceso de ajuste de diversas distribuciones estadísticas a un conjunto de datos, permitiendo comparar las distribuciones ajustadas y seleccionar la que mejor se adapta a los datos. Esto se logra mediante la implementación de criterios de comparación estadística como el AIC (Criterio de Información de Akaike) y el BIC (Criterio de Información Bayesiano).

Funciones Clave

- **Selección de Distribuciones**: La clase permite seleccionar un subconjunto de distribuciones disponibles en `scipy.stats` basado en el dominio de los datos y la naturaleza de las distribuciones (continuas o discretas). Esta selección puede ser personalizada según las necesidades específicas del análisis.
- **Ajuste de Distribuciones**: La clase ajusta múltiples distribuciones teóricas a los datos observados utilizando los métodos de máxima verosimilitud proporcionados por `scipy.stats`. Esto incluye la estimación de los parámetros de cada distribución.
- **Comparación de Distribuciones**: Después de ajustar las distribuciones, la clase calcula y compara los valores de log-verosimilitud, AIC y BIC para cada distribución. Estos criterios permiten evaluar la calidad del ajuste y seleccionar la distribución que mejor describe los datos.
- **Visualización de Resultados**: La clase incluye métodos para generar gráficos que superponen las curvas de densidad de las distribuciones ajustadas con el histograma de los datos. Esto facilita la interpretación visual de la calidad del ajuste.

El propósito principal de la clase `AnalisisDistribuciones` es proporcionar una manera sistemática y eficiente de ajustar múltiples distribuciones estadísticas a un conjunto de datos y evaluar cuál de estas distribuciones ofrece el mejor ajuste. Esto se realiza utilizando el test de Kolmogorov-Smirnov (KS) como criterio de evaluación principal.

Funciones Clave

- **Inicialización de Datos**: La clase se inicializa con un conjunto de datos que se desea analizar. Estos datos son utilizados para ajustar diversas distribuciones estadísticas continuas.
- **Lista de Distribuciones**: La clase incluye una lista de distribuciones estadísticas continuas disponibles en `scipy.stats`, tales como `norm`, `expon`, `uniform`, `lognorm`, `weibull_min`, `gamma`, `beta`, `chi2` y `t`.
- **Ajuste de Distribuciones**: El método `ajustar_distribuciones` ajusta cada una de las distribuciones continuas a los datos proporcionados. Para cada distribución, se calculan los parámetros de ajuste y se realiza el test de KS para evaluar la bondad del ajuste.
- **Evaluación de Resultados**: Los resultados del ajuste, incluyendo los valores del estadístico KS y su valor p, se almacenan en un diccionario que puede ser accedido para revisar el desempeño de cada distribución.
- **Selección de Mejores Distribuciones**: El método `obtener_mejores_distribuciones` permite seleccionar las mejores distribuciones basadas en un criterio específico, como el valor p del test de KS. Esto ayuda a identificar qué distribuciones se ajustan mejor a los datos.
- **Visualización de Resultados**: La clase incluye un método para graficar el histograma de los datos junto con las distribuciones mejor ajustadas. Esto facilita la interpretación visual de la calidad del ajuste.
- **Impresión de Resultados**: El método `imprimir_resultados` imprime los resultados del test de KS para todas las distribuciones consideradas, proporcionando una visión detallada del desempeño de cada distribución.

4 Imputación de Datos

La imputación de datos es un proceso fundamental en el análisis de datos, especialmente en el campo actuarial, donde la integridad y precisión de los datos son críticas para la evaluación de riesgos, la fijación de precios de seguros y la planificación financiera. Consiste en llenar los valores faltantes en un conjunto de datos con estimaciones razonables basadas en la información disponible. Veamos un ejemplo en nuestro codigo:

Tenemos una distribución Gamma original de 1000 datos, vamos a anular 17 valores luego procedemos a imputarlos.
Si los imputamos con el m]áximo:

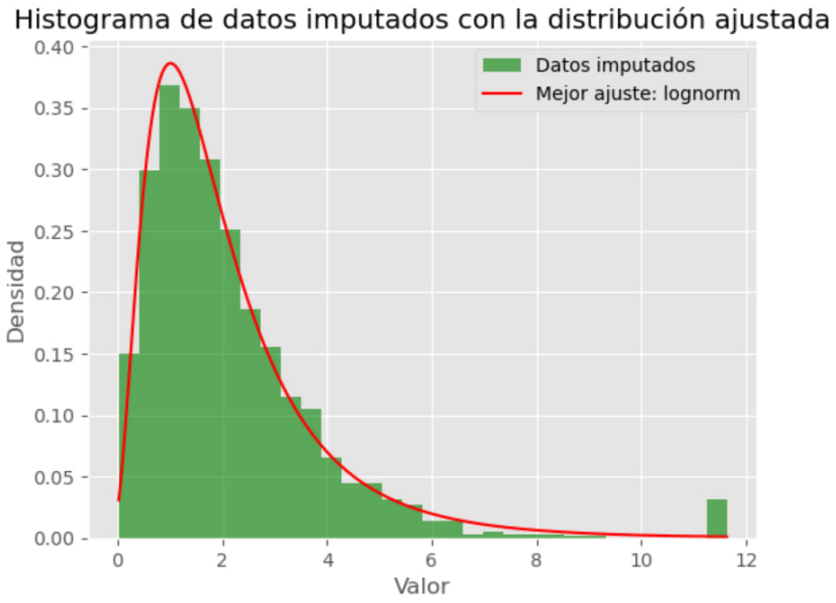

Figure 1: Imputación de 17 valores con el máximo

Note que en este caso nuestra distribución cambia a una lognormal.
Ahora, imputemos con la media:

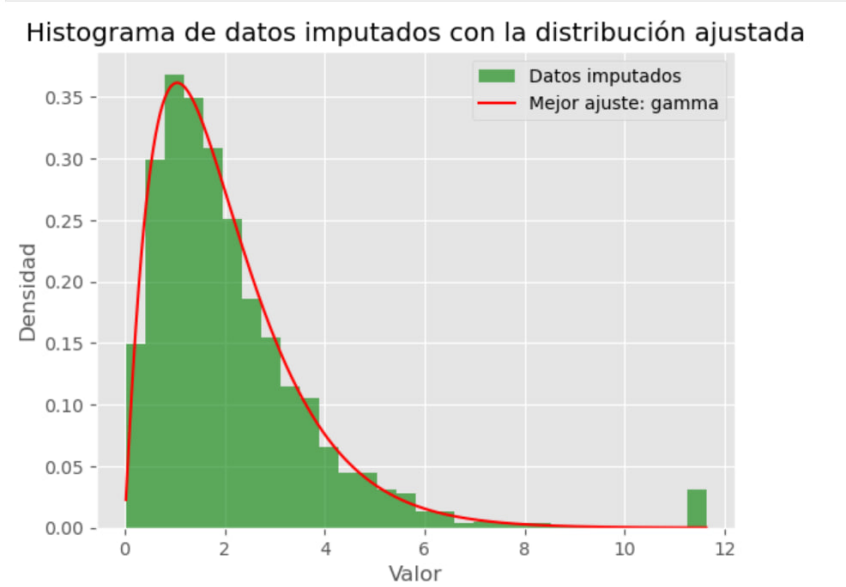

Figure 2: Imputación de 17 valores con la media

Si imputamos con la media seguimos teniendo nuestra distribucón original Gamma.

5 Conclusión

La selección de la distribución correcta es una parte esencial del análisis actuarial. Utilizar métodos como el AIC, el BIC y la prueba de Kolmogorov-Smirnov permite una evaluación exhaustiva de la adecuación de diferentes distribuciones a los datos. Estos métodos proporcionan herramientas valiosas para balancear la complejidad del modelo y la precisión del ajuste, garantizando decisiones basadas en datos robustos y confiables. En el contexto actuarial, donde la precisión en la modelización de riesgos es crítica, el uso adecuado de estos métodos contribuye significativamente a la eficacia y fiabilidad de las evaluaciones actuariales.

La imputación de valores faltantes es crucial en el análisis de datos, especialmente en contextos como el trabajo actuarial donde la precisión y la fiabilidad de los resultados son fundamentales. La elección del método adecuado de imputación puede afectar significativamente la calidad de los análisis y las decisiones basadas en ellos.

La coherencia en el tipo de distribución de los datos antes y después de la imputación es crucial para mantener la integridad de los análisis y las conclusiones derivadas. Cambiar inadvertidamente la distribución de los datos mediante métodos inapropiados de imputación puede llevar a interpretaciones erróneas.

Por lo tanto, seleccionar el método correcto de imputación depende de entender la distribución de los datos y elegir una estrategia que preserve o modele esta distribución de manera precisa. Esto no solo mejora la precisión de los resultados, sino que también fortalece la confianza en las conclusiones derivadas del análisis actuarial.