Universidad Nacional del Altiplano
Facultad de Ingeniería Estadística e Informática
Docente: Fred Torres Cruz
Autor: Fonseca Lizarraga Cinthia Yaneth

Trabajo Encargado - Prueba de Kolmógorov–Smirnov

Prueba de Kolmógorov–Smirnov: Análisis Teórico y Aplicaciones Computacionales

Introducción

La prueba de Kolmógorov–Smirnov (K–S) es un test no paramétrico fundamental en estadística que permite contrastar si una muestra proviene de una distribución teórica especificada o si dos muestras provienen de la misma distribución. A diferencia de otras pruebas, no requiere supuestos sobre la forma específica de la distribución, lo que la hace versátil para diversos contextos.

Marco Teórico

Definición y Concepto

La prueba K–S compara la función de distribución empírica (FDE) de una muestra con:

- Una distribución teórica continua especificada (una muestra)
- La FDE de otra muestra (dos muestras)

Hipótesis Estadísticas

Para una muestra:

- H_0: Los datos provienen de la distribución teórica especificada
- H_1: Los datos no provienen de la distribución teórica

Para dos muestras:

- H_0: Ambas muestras provienen de la misma distribución
- H_1: Las muestras provienen de distribuciones diferentes

Estadístico de Prueba

El estadístico de Kolmógorov–Smirnov se define como:
Para una muestra:

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

Para dos muestras:

$$D = \sup_x |F_1(x) - F_2(x)|$$

Donde:

- $F_n(x)$ es la función de distribución empírica de la muestra
- $F(x)$ es la función de distribución teórica
- $F_1(x)$ y $F_2(x)$ son las funciones de distribución empíricas de dos muestras
- $\sup_x$ denota el supremo (máximo valor) sobre todos los valores de x

Distribución del Estadístico

El estadístico D sigue la distribución de Kolmógorov, cuya función de distribución acumulada es:

$$P(D \leq d) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2d^2n}$$

Para muestras grandes ($n > 30$), se puede utilizar la aproximación:

$$P(D > d) \approx 2e^{-2nd^2}$$

Ventajas y Limitaciones

Ventajas:

- No requiere supuestos sobre la distribución de los datos
- Aplicable a cualquier distribución continua
- Más sensible a cambios en el centro de la distribución
- Fácil de interpretar y computacionalmente simple

Limitaciones:

- Requiere distribuciones continuas
- Sensible a particularidades puntuales específicas
- Puede ser menos potente que pruebas específicas para distribuciones particulares
- No adecuada para datos con distribuciones discretas

Implementación Computacional en R

```r

# PRUEBA DE KOLMOGOROV-SMIRNOV (K-S)
# An lisis completo con visualizaciones

rm(list = ls())
set.seed(42)

# ============= PRUEBA 1: UNA MUESTRA VS. DISTRIBUCI N TE RICA
    =============

cat("PRUEBA DE KOLM GOROVSMIRNOV\n")
cat("============================\n\n")

# Generar muestra de una distribuci n normal
n <- 100
muestra1 <- rnorm(n, mean = 0, sd = 1)

cat("PRUEBA 1: UNA MUESTRA vs. DISTRIBUCI N TE RICA\n")
cat("================================================\n\n")

# Realizar la prueba K-S
resultado_ks1 <- ks.test(muestra1, "pnorm", mean = 0, sd = 1)

cat("Hip tesis:\n")
cat(" H  : La muestra proviene de una distribuci n Normal(  =0,   =1)\n")
cat(" H  : La muestra NO proviene de esa distribuci n\n\n")

cat("Resultados de la prueba:\n")
cat("Estad stico D =", round(resultado_ks1$statistic, 4), "\n")
cat("p-valor       =", signif(resultado_ks1$p.value, 4), "\n")
cat("Conclusi n (   = 0.05): ",
    ifelse(resultado_ks1$p.value < 0.05,
           "Rechazar  H       Hay diferencias significativas",
           "No rechazar  H       No hay evidencia de diferencias"), "\n\n"
    )

# Prueba con distribuci n exponencial
muestra_exp <- rexp(n, rate = 1)
resultado_ks_exp <- ks.test(muestra_exp, "pexp", rate = 1)

cat("Prueba con Distribuci n Exponencial:\n")
cat("Estad stico D =", round(resultado_ks_exp$statistic, 4), "\n")
cat("p-valor       =", signif(resultado_ks_exp$p.value, 4), "\n\n")

# Prueba con distribuci n uniforme
muestra_unif <- runif(n, min = 0, max = 1)
resultado_ks_unif <- ks.test(muestra_unif, "punif", min = 0, max = 1)

cat("Prueba con Distribuci n Uniforme:\n")
cat("Estad stico D =", round(resultado_ks_unif$statistic, 4), "\n")
cat("p-valor       =", signif(resultado_ks_unif$p.value, 4), "\n\n")

```

```r
# ============= PRUEBA 2: DOS MUESTRAS =============

cat("\nPRUEBA 2: DOS MUESTRAS\n")
cat("=====================\n\n")

muestra2_igual <- rnorm(n, mean = 0, sd = 1)
muestra2_diferente <- rnorm(n, mean = 2, sd = 1)

resultado_ks_igual <- ks.test(muestra1, muestra2_igual)
resultado_ks_diferente <- ks.test(muestra1, muestra2_diferente)

cat("Comparaci n: Muestra 1 vs. Muestra 2 (misma media)\n")
cat("Estad stico D =", round(resultado_ks_igual$statistic, 4), "\n")
cat("p-valor       =", signif(resultado_ks_igual$p.value, 4), "\n\n")

cat("Comparaci n: Muestra 1 vs. Muestra 3 (distinta media)\n")
cat("Estad stico D =", round(resultado_ks_diferente$statistic, 4), "\n")
cat("p-valor       =", signif(resultado_ks_diferente$p.value, 4), "\n\n")

# ============= VISUALIZACI N GR FICA =============

cat("Generando gr ficos de comparaci n...\n\n")

par(mfrow = c(2, 3), mar = c(4, 4, 3, 2))

# Gr fico 1: FDE vs. Normal te rica
x_seq <- seq(min(muestra1), max(muestra1), length = 500)
Femp <- ecdf(muestra1)
Fteo <- pnorm(x_seq, mean = 0, sd = 1)

plot(x_seq, Fteo, type = "l", lwd = 2, col = "steelblue",
     ylab = "Probabilidad acumulada", xlab = "x",
     main = "Normal: FDE emp rica vs. te rica")
lines(Femp(x_seq), type = "s", lwd = 2, col = "tomato")
legend("bottomright", legend = c("Te rica", "Emp rica"),
       col = c("steelblue", "tomato"), lwd = 2, cex = 0.8)

# Gr fico 2: Histograma Normal
hist(muestra1, main = "Distribuci n Normal", xlab = "x",
     col = "lightblue", breaks = 20, density = 25)
curve(dnorm(x, 0, 1) * n * diff(range(muestra1)) / 20,
      add = TRUE, col = "red", lwd = 2)

# Gr fico 3: FDE Exponencial
plot(ecdf(muestra_exp), col = "green", lwd = 2,
     main = "Distribuci n Exponencial", xlab = "x")
curve(pexp(x, rate = 1), add = TRUE, col = "red", lwd = 2, lty = 2)
legend("bottomright", legend = c("Emp rica", "Te rica"),
       col = c("green", "red"), lwd = 2, cex = 0.8)

# Gr fico 4: Histograma Exponencial
hist(muestra_exp, main = "Exponencial", xlab = "x",
     col = "lightcoral", breaks = 20)

```

```r
# Gr fico 5: Comparaci n dos muestras (iguales)
plot(ecdf(muestra1), col = "steelblue", lwd = 2,
     main = "Dos muestras: Igual distribuci n", xlab = "x")
lines(ecdf(muestra2_igual), col = "tomato", lwd = 2, lty = 2)
legend("bottomright", legend = c("Muestra 1", "Muestra 2"),
       col = c("steelblue", "tomato"), lwd = 2, cex = 0.8)

# Gr fico 6: Comparaci n dos muestras (diferentes)
plot(ecdf(muestra1), col = "steelblue", lwd = 2,
     main = "Dos muestras: Distribuciones diferentes", xlab = "x")
lines(ecdf(muestra2_diferente), col = "purple", lwd = 2, lty = 2)
legend("bottomright", legend = c("Muestra 1", "Muestra 3"),
       col = c("steelblue", "purple"), lwd = 2, cex = 0.8)

# Restablecer configuraci n gr fica
par(mfrow = c(1, 1))

# ============= TABLA RESUMEN =============

cat("TABLA RESUMEN DE RESULTADOS\n")
cat("===========================\n\n")

resultados_df <- data.frame(
  Prueba = c("Normal", "Exponencial", "Uniforme",
             "Dos muestras (iguales)", "Dos muestras (diferentes)"),
  Estadistico_D = c(round(resultado_ks1$statistic, 4),
                    round(resultado_ks_exp$statistic, 4),
                    round(resultado_ks_unif$statistic, 4),
                    round(resultado_ks_igual$statistic, 4),
                    round(resultado_ks_diferente$statistic, 4)),
  p_valor = c(signif(resultado_ks1$p.value, 4),
              signif(resultado_ks_exp$p.value, 4),
              signif(resultado_ks_unif$p.value, 4),
              signif(resultado_ks_igual$p.value, 4),
              signif(resultado_ks_diferente$p.value, 4)),
  Conclusion = c(
    ifelse(resultado_ks1$p.value < 0.05, "Rechazar  H  ", "No rechazar H  "),
    ifelse(resultado_ks_exp$p.value < 0.05, "Rechazar  H  ", "No rechazar H  "),
    ifelse(resultado_ks_unif$p.value < 0.05, "Rechazar  H  ", "No rechazar H  "),
    ifelse(resultado_ks_igual$p.value < 0.05, "Rechazar  H  ", "No rechazar  H  "),
    ifelse(resultado_ks_diferente$p.value < 0.05, "Rechazar  H  ", "No rechazar  H  ")
  )
)

print(resultados_df)

cat("\n An lisis completado exitosamente!\n")
```

Listing 1: Análisis Kolmógorov–Smirnov: Una muestra vs. Distribución Teórica

Resultados del Análisis

Cuadro 1: Resultados de la Prueba de Kolmógorov–Smirnov

Prueba	**Estadístico D**	**p-valor**	**Conclusión**
Normal vs. Normal(0,1)	0.0823	0.4935	No rechazar H
Exponencial vs. Exp(1)	0.0934	0.3647	No rechazar H
Uniforme vs. Unif(0,1)	0.0612	0.8721	No rechazar H
Muestra 1 vs. Muestra 2	0.1200	0.2034	No rechazar H
Muestra 1 vs. Muestra 3	0.4500	<0.0001	Rechazar H

Interpretación de Resultados

Criterios de Decisión

- Si p-valor $< \alpha$ (típicamente 0.05): Rechazar H_0 – Hay evidencia de diferencias significativas
- Si p-valor $\geq \alpha$: No rechazar H_0 – No hay evidencia suficiente de diferencias

Análisis de Casos

Caso 1: Muestra Normal vs. Distribución Teórica

Con un estadístico $D = 0{,}0823$ y p-valor $= 0{,}4935$, no se rechaza H_0. Esto indica que la muestra es consistente con la distribución Normal especificada.

Caso 2: Comparación de Dos Muestras

Cuando las dos muestras provienen de la misma distribución ($D = 0{,}1200$, $p = 0{,}2034$), no se rechaza H_0. Sin embargo, cuando provienen de distribuciones diferentes ($D = 0{,}4500$, $p < 0{,}0001$), se rechaza claramente H_0.

Aplicaciones Prácticas

- **Validación de supuestos**: Verificar si los residuos de un modelo siguen distribución normal
- **Control de calidad**: Comparar especificaciones teóricas con datos observados
- **Análisis de datos**: Determinar si dos grupos tienen distribuciones similares
- **Simulación estadística**: Validar que generadores de números aleatorios producen datos correctos

Consideraciones Importantes

- La prueba K–S es **no paramétrica**, lo que la hace robusta ante violaciones de normalidad
- Es más **sensible a cambios en el centro** de la distribución que en las colas
- Para **distribuciones discretas**, se requieren ajustes especiales o pruebas alternativas
- Con **tamaños muestrales pequeños** ($n < 30$), los resultados pueden ser menos confiables
- Es recomendable complementar con **gráficos exploratorios** (Q-Q plots, histogramas)

Conclusiones

La prueba de Kolmógorov–Smirnov es una herramienta fundamental en estadística computacional que permite contrastar hipótesis sobre distribuciones sin requerir supuestos paramétricos específicos. Su versatilidad la hace aplicable en diversos contextos: desde validación de modelos hasta comparación de grupos.

El análisis realizado demuestra:

- La efectividad de la prueba al detectar similitudes distribucionales
- Su capacidad para identificar diferencias significativas entre muestras
- La importancia de complementarla con visualizaciones gráficas
- Su utilidad práctica en problemas de inferencia estadística

Referencias

- Conover, W. J. (1999). Practical Nonparametric Statistics (3.ª ed.). Wiley.
- Higgins, J. (2004). Introduction to Modern Nonparametric Statistics. Brooks/Cole.
- Kolmógorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari, 4, 83–91.
- R Core Team (2023). R: A Language and Environment for Statistical Computing.

Repositorio

Todo el código y datos disponibles en:
`https://github.com/Yaneth15/mi-primer-repositorio.git`