ANÁLISIS ESTADÍSTICO DE TENDENCIAS

INDER TECUAPETLA

1. Prueba de Mann-Kendall

A continuación discutiremos la prueba Mann-Kendall, una prueba estadística *no paramétrica* para determinar cuantitativamente la presencia de tendencia en series de tiempo. Los siguientes párrafos están basados en Mann (1945) y en los capítulos 4 y 5 de Kendall (1962).

Supongamos que tenemos elementos aleatorios $X_1, \ldots, X_n$ que son independientes y cada uno posee la misma distribución de probabilidad. No es necesario suponer que la distribución de las X_is es normal. Parafraseando a Kendall supondremos que *no existe alguna relación* entre estas muestras poblacionales (X_is). Parafraseando a Mann supondremos que la muestra (todo el conjunto de las X_is) es completamente aleatoria. Para simplificar la presentación supondremos que $X_i \neq X_j$ para toda j e i.

Consideremos el estadístico

$$S_n = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{signo}(X_j - X_i), \tag{1}$$

donde

$$\mathrm{signo}(X_j - X_i) = \begin{cases} +1 & \text{si } X_j - X_i > 0 \\ -1 & \text{si } X_j - X_i < 0 \end{cases}.$$

Siguiendo la presentación de Mann es fácil demostrar que el valor medio de S_n es cero. Este hecho abre la puerta a la utilización del estadístico S_n para determinar tendencia en una serie de tiempo.

En efecto, cuando calculamos $\widehat{S}_n$, la contraparte muestral de S_n, bajo el supuesto de ausencia de relación entre las observaciones o no tendencia, entonces esperamos que $\widehat{S}_n$ sea pequeño, cercano a cero. Por otra parte, si $|\widehat{S}_n|$ resulta *muy* grande entonces es cuestionable el supuesto de que no exise relación entre las observaciones. Nota que $|\widehat{S}_n|$ es muy grande cuando $\widehat{S}_n$ es un número positivo grande o cuando $\widehat{S}_n$ es un número negativo grande. Observa que el primero de estos casos ocurre cuando en (1) las observaciones más recientes dominan a las observaciones más tempranas; este caso está en correspondencia con una tendencia creciente o positiva. Similarmente, cuando las observaciones tempranas dominan a las más recientes entonces $\widehat{S}_n$ es negativo y grande; este caso se asocia a la tendencia negativa o decreciente en una serie de tiempo.

Como en toda prueba de hipótesis, el cálculo del p-valor depende de la hipótesis de interés. Consideramos que los ejemplos de abajo aclararán la mecánica del cálculo de esta probabilidad.

Date: 6/24/2022

itecuapetla@conabio.gob.mx.

Antes de pasar a los ejemplos notamos que la distribución de probabilidades de S_n es complicada de establecer; existen tablas para $n \leq 10$. Para valores grandes de n, sin embargo, hacemos uso de la siguiente aproximación normal.[1]

Arriba mencionamos que el primer momento central de S_n es cero. Algo de aritmética y algunas propiedades del funcional valor esperado nos permiten ver que

$$\mathsf{VAR}(S_n) = \frac{n(n-1)(2n+5)}{18} =: \sigma_n$$

Usando elementos elegantes, uno de análisis y el otro de combinatoria, Mann y Kendall establecieron que cuando n crece inconmensurablemente entonces los momentos centrales de S_n son iguales a los momentos centrales de una normal con media cero y varianza σ_n. Los detalles de esta demostración se encuentran en las referencias dadas arriba.

La Figura 1 muestra la aproximación descrita arriba para algunos valores de n. Como se nota en esta figura S_n toma valores en los números enteros, su distribución es simétrica alrededor de cero y no es difícil ver que sus valores máximo y mínimo son $n(n-1)/2$ y $-n(n-1)/2$.[2] Considerando estos elementos, el estadístico para probar tendencia tiene la siguiente forma

$$T_n = \begin{cases} \frac{S_n - 1}{\sqrt{\mathsf{VAR}(S_n)}} & \text{si } S_n > 0 \\ \frac{S_n + 1}{\sqrt{\mathsf{VAR}(S_n)}} & \text{si } S_n < 0 \end{cases}.$$

La resta y suma de 1 en la definición de T_n obedece a una conocida *corrección por continuidad*; esto mejora la precisión en el cálculo de probabilidades al usar la aproximación normal.

1.1. Ejemplos.

1.1.1. Ruido blanco. El término **ruido blanco** se usa para describir procesos aleatorios con media cero, varianza constante (y típicamente igual a uno) e independientes. Podemos usar R para simular un ruido blanco. En efecto, con `rnorm(50)` simulamos una muestra de números aleatorios gausianos con media cero, varianza uno e independientes. Por tanto en esta muestra, y en el ruido blanco en general, no existe relación de dependencia entre las observaciones. Por tanto la hipótesis nula es que *no existe asociación* entre las observaciones y la hipótesis alternativa es *que existe algún tipo de asociación.*

Como mencionamos arriba, estas hipótesis pueden interpretarse como H_0 : no existe tendencia y H_a : existe tendencia (positiva o negativa). Por tanto el p-valor en este caso es igual a

$$\mathbf{P}\{|T_n| > |\widehat{T}_n|\} = 2\,\mathbf{P}\{T_n > |\widehat{T}_n|\}.$$

Acá el código de esta simulación y la aplicación de la prueba Mann-Kendall.

```
muestra <- rnorm(50)
(out <- mk.test(muestra))
```

```
##
##  Mann-Kendall trend test
##
## data:  muestra
```

[1]Sí, vamos a suponer que un valor grande es $n > 10$.

[2]Nota que si $X_j > X_i$ para toda j e i, entonces S_n es equivalente a sumar $n - 1 + n - 2 + \cdots + 2 + 1$.

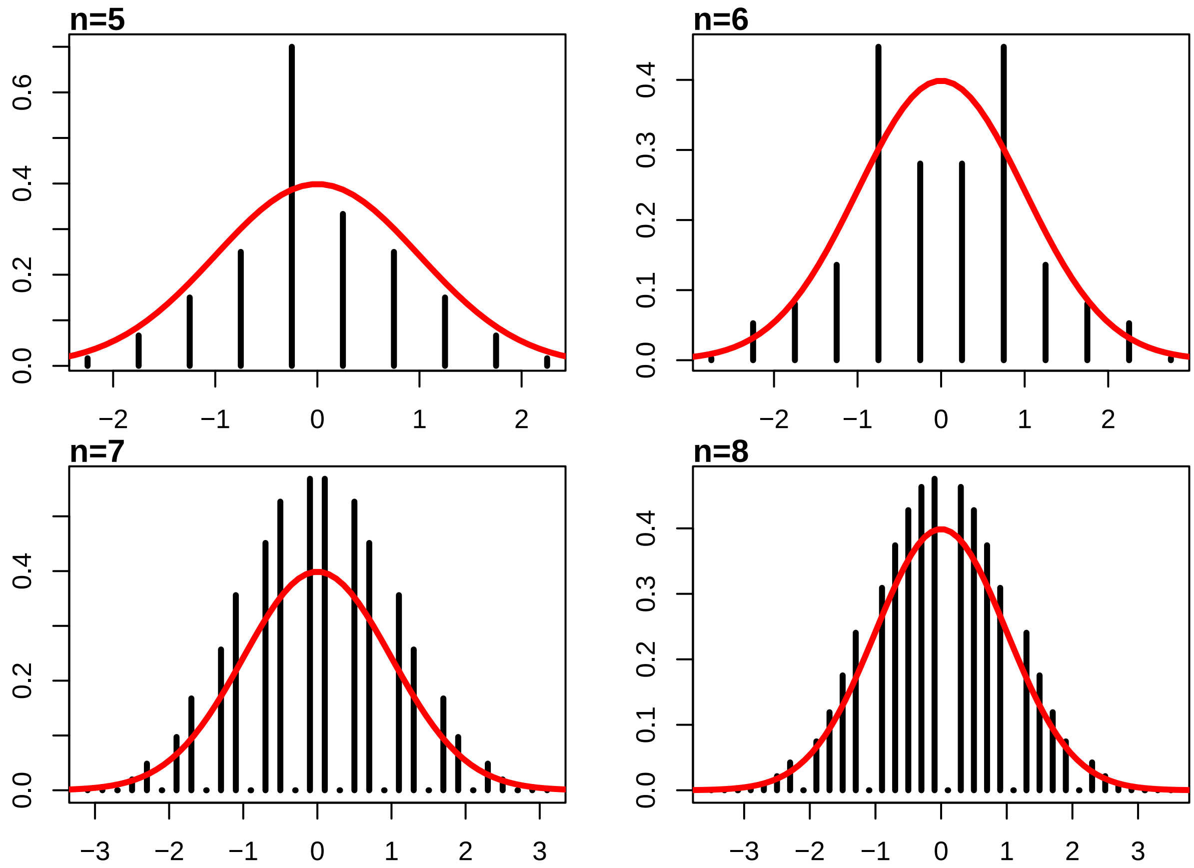

FIGURA 1. Aproximación normal del estadístico Sn. La línea roja muestra la densidad de una normal estándar. Las barras negras muestran el histograma de distribución de Sn.

```
## z = 0.70265, n = 50, p-value = 0.4823
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 8.500000e+01 1.429167e+04 6.938776e-02
```

Nota que el objeto `out` es una lista con otros nueve objetos. Con `out$estimates` podemos calcular el p-valor, y por completitud, y compararlo con el calculado por la función `mk.test`:

```
str(out)
```

```
## List of 9
##  $ data.name  : chr "muestra"
##  $ p.value    : num 0.482
##  $ statistic  : Named num 0.703
##   ..- attr(*, "names")= chr "z"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "S"
##  $ parameter  : Named int 50
##   ..- attr(*, "names")= chr "n"
```

```
##  $ estimates  : Named num [1:3] 8.50e+01 1.43e+04 6.94e-02
##   ..- attr(*, "names")= chr [1:3] "S" "varS" "tau"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Mann-Kendall trend test"
##  $ pvalg      : num 0.482
##  - attr(*, "class")= chr "htest"
```

```
Tn <- (out$estimates[1]-1) / sqrt(out$estimates[2])
pVal <- 2 * (1 - pnorm(abs(as.numeric(Tn))))
pVal
```

```
## [1] 0.4822751
```

```
out$p.value
```

```
## [1] 0.4822751
```

Con un p-valor así de grande no podemos rechazar la hipótesis nula de no existencia de tendencia en las observaciones simuladas -lo cual es un resultado esperado.

1.1.2. Producción trimestral de cerveza. La Figura 2 muestra la producción trimestral de cerveza en Australia en el periodo 1960-1973 junto con una tendencia estimada (panel **A**); visualmente notamos un aparente crecimiento sostenido de la tendencia. Podemos agregar contenido cuantitativo a esta impresión a través de utilizar la prueba de Mann-Kendall en la tendencia estimada.

En específico, tenemos el interés de establecer que el promedio de la producción trimestral de cerveza en Australia ha aumentado. Así nuestra hipótesis nula es H_0 : no existe tendencia y la hipótesis alternativa en este caso es H_a : existe una tendencia positiva.

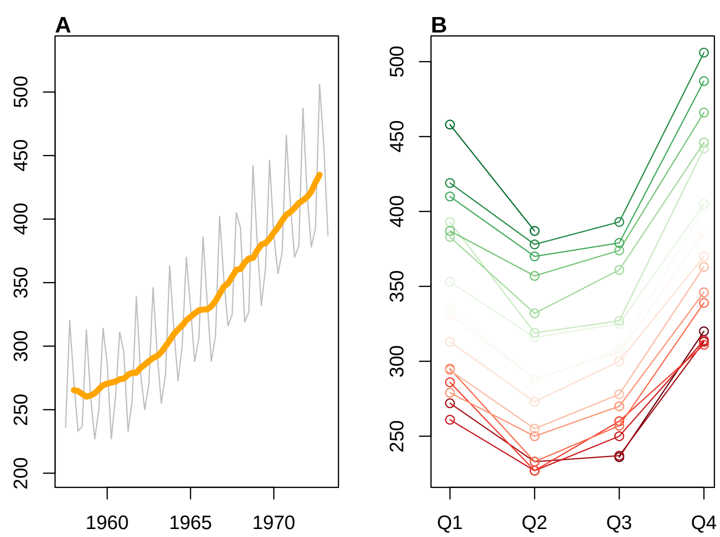

FIGURA 2. Producción trimestral de cerveza en Australia (megalitros) del 3er trimestre de 1957 al 2do trimestre de 1973.

De acuerdo a estas especificaciones el p-valor asociado a la hipótesis nula es igual a $\mathbf{P}\{T_n > \widehat{T}_n\}$. En R obtenemos un p-valor de 3.9×10^{-29} por lo cual rechazamos **contundentemente** la hipótesis nula (en favor de la alternativa). Con un p-valor tan bajo se concluye que en el periodo en consideración la producción media de cerveza en Australia observó un crecimiento.

```
mk <- mk.test(x = as.numeric(trendBeer[!is.na(trendBeer)]),
              alternative = "greater")
Tn <- (mk$estimates[1]-1) / sqrt(mk$estimates[2])
pVal <- 1 - pnorm(as.numeric(Tn))
pVal
```

```
## [1] 0
```

```
mk$p.value
```

```
## [1] 3.906445e-29
```

1.1.3. Concentración de sedimentos en el Río Rhine. El río Rhine es una de los más importantes de Europa; nace en Suiza y su flujo se mueve principalmente hacia el norte pasando por Alemania y los Países Bajos. En el paquete `trend` se encuentra la base de datos `maxau` la cual reporta series de tiempo anuales de la concentración promedio de sedimentos (en mg/l) y la descarga promedio (en m^3) desde 1965 hasta 2009. Aquí nos concentraremos en la concentración de sedimentos.

La Figura 3 muestra la serie de tiempo de la concentración de sedimentos junto con su tendencia estimada vía LOWESS (Cleveland (1979)); aparentemente, ha habido un decrecimiento en la concetración de sedimentos desde 1965 hasta 2009. Usamos la prueba Mann-Kendall para probar esta hipótesis cuantitativamente.

En este caso tenemos las hipótesis H_0 : no existe tendencia y H_a : existe tendencia negativa. El p-valor correspondiente a esta hipótesis es igual a $\mathbf{P}\{T_n \leq \widehat{T}_n\}$.

En R usamos el siguiente código para efectuar la prueba Mann-Kendall a la serie de tiempo de sedimentos:

```
out <- mk.test(sedimentos, alternative = "less")
out
```

```
##
##  Mann-Kendall trend test
##
## data:  sedimentos
## z = -3.8445, n = 45, p-value = 6.041e-05
## alternative hypothesis: true S is less than 0
## sample estimates:
##             S          varS            tau
##  -394.0000000 10450.0000000    -0.3979798
```

```
Tn <- (out$estimates[1]+1) / sqrt(out$estimates[2])

pVal <- pnorm(as.numeric(Tn))
pVal
```

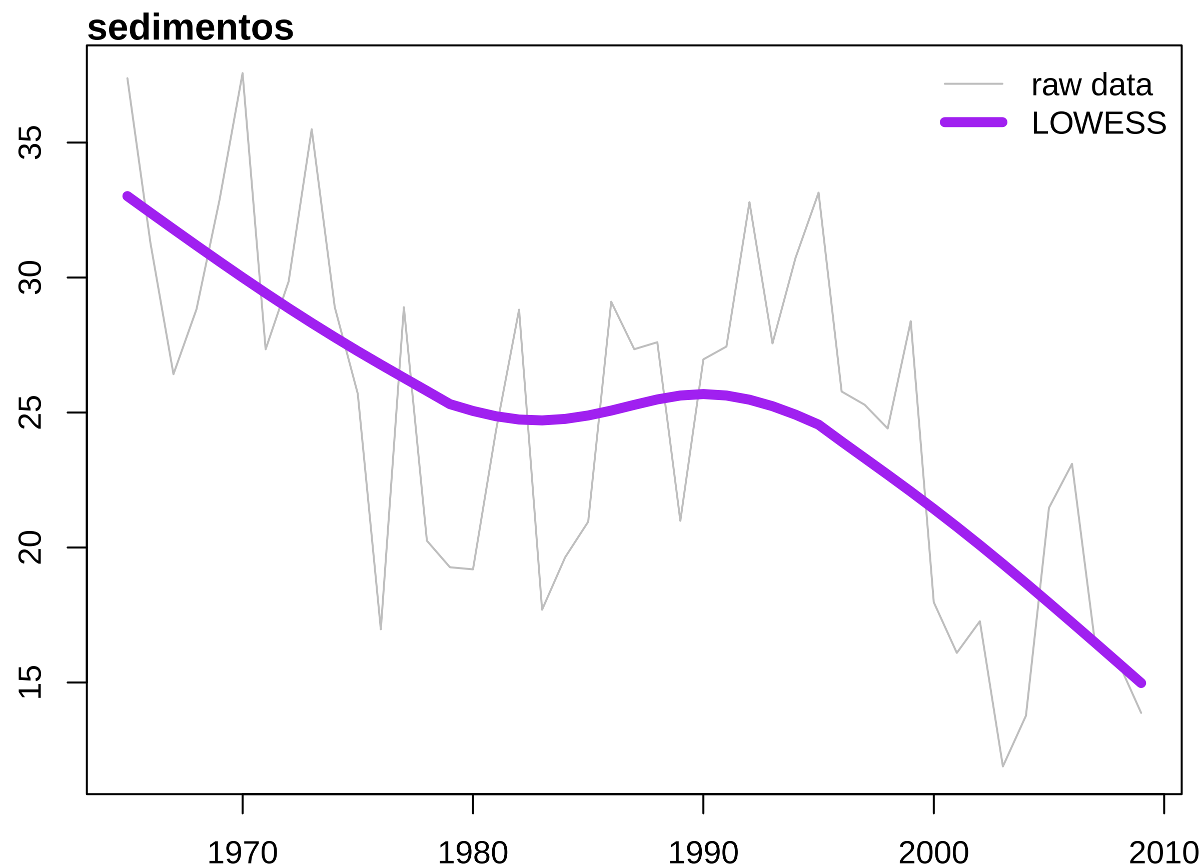

Figura 3. Concentración anual promedio de sedimentos en el Rhine de 1965 a 2009.

```
## [1] 6.041115e-05
```

```
out$p.value
```

```
## [1] 6.041115e-05
```

A la luz de estos datos y el correspondiente p-valor de la prueba, podemos concluir que existe evidencia estadística de un decrecimiento en la concentración promedio de sedimentos del río Rhine durante el periodo 1965-2009.

2. Estimador de tendencia lineal Theil-Sen

En esta breve sección mostramos una herramienta para estimar la aparente tendencia lineal en una serie de tiempo. Nota que la prueba Mann-Kendall ayuda a determinar la existencia de tendencia, no necesariamente lineal, en una serie de tiempo. Las siguientes ideas están basadas en Sen (1968).

Supongamos que tenemos elementos aleatorios $Y_1, \ldots, Y_n$ que son independientes y cada uno posee la misma distribución de probabilidad. No es necesario suponer que la distribución de las Y_is es normal. Para simplificar la presentación suponemos que $Y_i \neq Y_j$ para toda j e i.

Supongamos también que en cualquier punto del tiempo t, $Y_t = a + b\,t$. La propuesta de Sen es estimar a y b de modo robusto. Observa que la pendiente entre cualesquiera dos puntos

(i, Y_i) y (j, Y_j) está dado por

$$b_{i,j} = \frac{Y_j - Y_i}{j - i}.$$

Nota que $b_{i,j}$ es una cantidad aleatoria, ya que Y_i y Y_j son aleatorios, y consecuentemente el conjunto de pendientes $\{b_{i,j} : 1 \leq i < j \leq n\}$ tiene una distribución de probabilidades. Aunque esta distribución es desconocida, esto no nos impide calcular la mediana de esta distribución y utilizarla como un estimador robusto de b. Específicamente,

$$\widehat{b} = \text{mediana}\{b_{i,j} : 1 \leq i < j \leq n\}.$$

Similarmente, un estimador robusto para a está dado por

$$\widehat{a} = \text{mediana}\{Y_t - \widehat{b}t : 1 \leq t \leq n\}.$$

En la referencia dada arriba podemos encontrar algunas propiedades de los estimadores $\widehat{a}$ y $\widehat{b}$.

Concluimos esta sección usando el estimador de Theil-Sen para describir una tendencia lineal en la serie de tiempo de sedimentos del río Rhine. Abajo se muestra el código usado para generar la Figura 4.

```
out <- sens.slope(sedimentos)
out
```

```
##
##  Sen's slope
##
## data:  sedimentos
## z = -3.8445, n = 45, p-value = 0.0001208
## alternative hypothesis: true z is not equal to 0
## 95 percent confidence interval:
##  -0.4196477 -0.1519026
## sample estimates:
## Sen's slope
##  -0.2876139
```

```
b_hat <- as.numeric(out$estimates)
a_hat <- median( sedimentos - b_hat * 1:length(sedimentos) )

lineaTheilSen <- ts(a_hat + b_hat * 1:length(sedimentos), start = c(1965, 1),
                    end = c(2009, 1), frequency = 1)

par(mar = c(2,2,1,2), adj=0)
plot(sedimentos, col = "gray", ylab = "mg/l", main = "sedimentos")
lines(lineaTheilSen, lwd = 5, col = "lightcoral")
legend("topright", legend = c("raw data", "linear trend"),
       col = c("gray", "lightcoral"), lty = rep(1,2), lwd = c(1,5), bty = "n")
```

Bibliografía

Cleveland, William S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74 (368): 829–36.

Kendall, Maurice George. 1962. *Rank Correlation Methods*. 4th ed. Griffin.

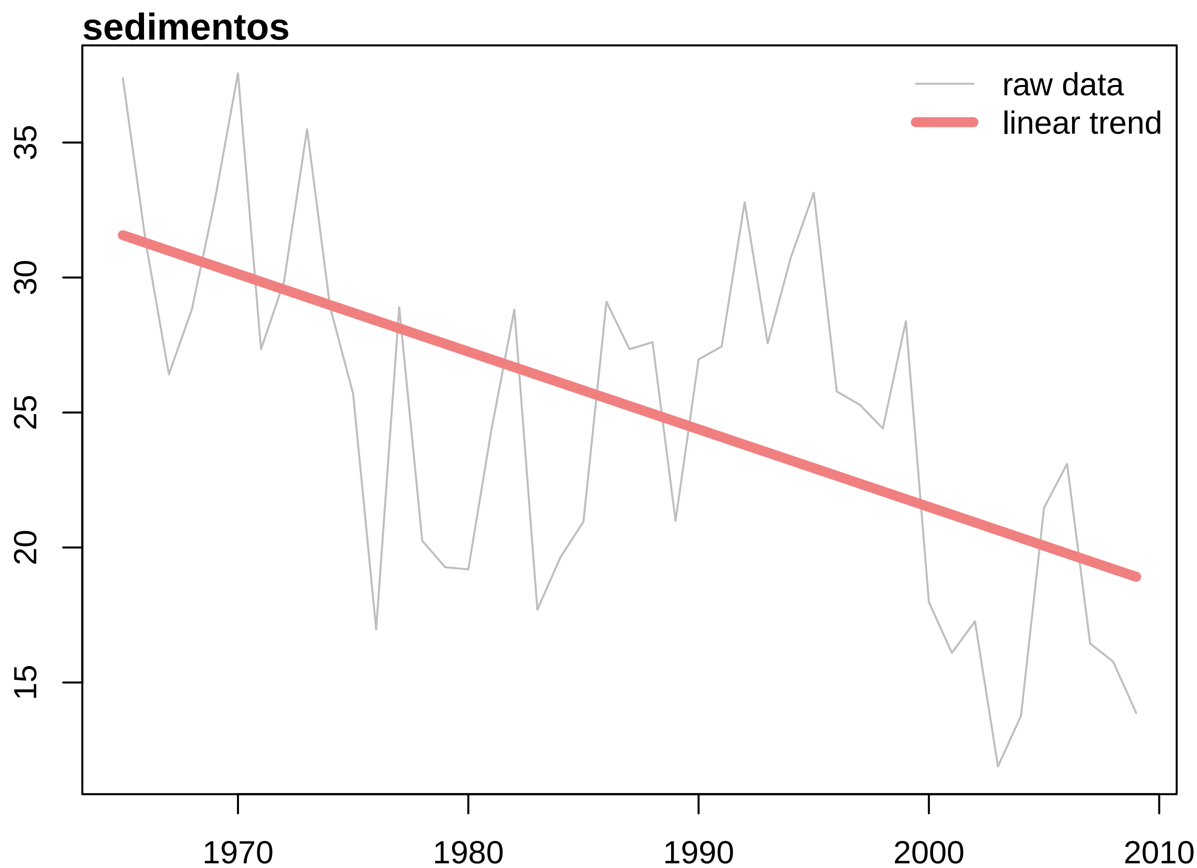

Figura 4. Estimador de tendencia lineal.

Mann, Henry B. 1945. "Nonparametric Tests Against Trend." *Econometrica: Journal of the Econometric Society*, 245–59.

Sen, Pranab Kumar. 1968. "Estimates of the Regression Coefficient Based on Kendall's Tau." *Journal of the American Statistical Association* 63 (324): 1379–89.