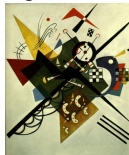


Simulación

Pruebas para números pseudoaleatorios.

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México



Semana 6

ITAM

Pruebas para números pseudoaleatorios.

Propiedades de los buenos generadores I



Recordemos algunas propiedades deseables en números pseudoaleatorios que se piden para tener calidad para propósitos de simulación:

1. Los números *deben* parecer distribuirse uniformemente y ser independientes.
2. Los métodos para generarlos deben ser rápidos y eficientes.
3. Deben ser capaces de replicarse.
4. Se debería poder generar más de una secuencia de números.
5. Un generador debe tener periodo muy largo.

Propiedades de los buenos generadores II

- El punto [2] se cumple utilizando, por ejemplo, GLC's.
- El punto [3] se puede cumplir en R, utilizando la función `set.seed(x)` o en python con `random.seed()` (`import random`). De esta forma se pueden generar las mismas secuencias de números aleatorios cuando se usa un generador.
- El punto [4] lo podemos poner en práctica utilizando un generador con periodo grande, y generado las diferentes secuencias con diferentes semillas. Por ejemplo, podemos generar la primer secuencia de longitud n_1 con cualquier valor, una segunda secuencia haciendo $Z_0 = Z_{n_1}$ de longitud n_2 y así sucesivamente para varios valores n_1, n_2, \dots, n_k .
- El punto [5] lo cumplimos con la adecuada selección de parámetros y utilizando generadores de periodo extendido.

Entonces sólo nos queda revisar el punto [1]: para eso revisaremos técnicas y métodos para verificar que los números se comportan aleatoriamente de manera uniforme y son independientes.

Tipos de pruebas para números pseudoaleatorios I

- Maurice Kendall y Babington-Smith introdujeron el concepto de **aleatoriedad local**: cada subsucesión 'razonablemente' larga, debe *parecer* aleatoria y pasar un conjunto simple de pruebas estadísticas.
- Ellos propusieron un conjunto o batería pequeña de pruebas:
 - 1 La frecuencia de cada dígito
 - 2 La frecuencia de cada par en valores sucesivos (prueba serial)
 - 3 La frecuencia de ciertos bloques de cinco dígitos (poker)
 - 4 longitud de los gaps entre las ocurrencias de un dígito dado

Las frecuencias son comparadas contra las esperadas via una prueba de bondad de ajuste basada en la distribución χ^2 . Replicaron las pruebas con partes disjuntas de su tabla de dígitos aleatorios, y pasaron las pruebas.

- Siguiendo la misma idea, se han definido **baterías de pruebas** estadísticas de distinta complejidad tales que conjuntos de números pseudialeatorios a prueba tienen que pasar.
- En general, se consideran dos tipos de pruebas:

Tipos de pruebas para números pseudoaleatorios II

- **Empíricas:** se basan en muestras de los números y se aplican tanto pruebas de bondad de ajuste como de independencia.
- **Teóricas:** se basan en las características numéricas de los generadores. Usualmente son muy sofisticadas y de carácter global. No usan muestras sino teoremas matemáticos.
- Nos concentraremos en las pruebas empíricas.

Tipo de prueba:	Bondad de ajuste	Independencia ¹
Hipótesis a probar	$H_0 : u_i \sim F$	$H_0 : u_i \perp\!\!\!\perp u_j \forall i \neq j$
Ejemplos	Kolmogorov-Smirnov (KS). Pruebas de bondad de ajuste χ^2 Anderson-Darling, Cramér-von Mises qq-plots. (no es formal)	Rachas (número y longitud) Autocorrelación gaps o espacios póker

- Existen muchas más pruebas para cada caso. Aquí la imaginación es el límite. Nos concentraremos en las pruebas que se indican en la tabla anterior.
- Al finalizar esta sección se comentarán otros conjuntos de pruebas para números aleatorios.

¹el símbolo $\perp\!\!\!\perp$ significa independencia entre variables aleatorias

¿Cuándo se aplican las pruebas? I

- Las pruebas se aplican cuando:
 - No se conoce el método utilizado para generar los números aleatorios.
 - Se está experimentando con un nuevo generador.
 - El método utilizado no está bien documentado.
 - Se mezclan métodos en una simulación muy grande.
 - Se quiere verificar que no hay errores de programación en el generador.
- Sin embargo, *aún si un conjunto de números pasa todas las pruebas, no hay garantía absoluta de aleatoriedad.*
 - Se debe considerar la aplicación de múltiples pruebas a una muestra. Tomar en cuenta entonces la desigualdad de Bonferroni.
 - Cuando una misma prueba se aplica a **varias submuestras** de números del generador bajo observación, tomar en cuenta que se puede rechazar hasta el error tipo I (α).
- Las pruebas se pueden utilizar para probar que una muestra viene de alguna distribución específica.
 - Cualquier dato de entrada a un modelo de simulación puede probarse contra la distribución objetivo.

¿Cuándo se aplican las pruebas? II

- En la práctica, determinar la distribución apropiada para los datos de entrada en una simulación es una tarea a la que hay que dedicar tiempo y consume recursos.
- Proceso equivalente al proceso de 'data cleaning' en análisis de datos.
- Hay 4 pasos en el desarrollo de un modelo útil para datos de entrada para una simulación:
 - 1 Recabar datos del sistema real de interés.
 - 2 Identificar una distribución de probabilidad que represente el proceso de entrada.
 - 3 Estimar los parámetros adecuados para el modelo de probabilidad correspondiente (si son modelos paramétricos).
 - 4 Evaluar la distribución y parámetros escogidos para bondad de ajuste.

Pruebas de bondad de ajuste

Función de distribución empírica

La prueba de Kolmogorov-Smirnov, es una prueba de bondad de ajuste para funciones de distribución basada en la *distribución empírica*. Esta distribución es la base para un conjunto de pruebas conocidas como *no paramétricas* o *libres de distribución*.

Definición. Función de distribución empírica (EDF)

Dada una muestra aleatoria X_1, \dots, X_n de variables con función de distribución F , se define como la función:

$$F_n(x) = \frac{\#(X_i \leq x)}{n} = \frac{\sum_{i=1}^n I_{(-\infty, x]}(X_i)}{n}$$

- Ejercicio 1: para la muestra aleatoria 1,0,1,1,0,0,0,1,0,1,0, calcular su EDF
- Ejercicio 2: para la muestra aleatoria 7, 9, 6, 1, 7, 9, 4, 7, 3, 5, 4, 5, 5, 6, 9, calcular su EDF

Función de distribución empírica I

Hay una estrecha relación entre la distribución empírica y las *estadísticas de orden*.

Definición. Estadísticas de orden

Si X_1, \dots, X_n es una muestra aleatoria de una distribución F , las *estadísticas de orden* se definen como los datos ordenados de menor a mayor: $X_{(1)} \leq \dots \leq X_{(n)}$, donde:

- $X_{(1)} = \text{mín}\{X_1, \dots, X_n\}$
- ...
- $X_{(n)} = \text{máx}\{X_1, \dots, X_n\}$

En particular, la función de distribución tiene saltos en los valores de las estadísticas de orden:

$$F_n(x) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ i/n & \text{si } X_{(i-1)} \leq x < X_{(i)}, \quad i = 1, \dots, n \\ 1 & \text{si } x \geq X_{(n)} \end{cases}$$

donde $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ son las estadísticas de orden asociadas a la muestra.

Función de distribución empírica II

Cuando hay empates entre los valores, los escalones de $F_n(x)$ son del tamaño del número de valores repetidos de $X_{(j)}$.

Distribución de la función de distribución empírica

Teorema

Sea F_n la función de distribución empírica para una muestra aleatoria X_1, \dots, X_n de F . Entonces

$$P\left(F_n(x) = \frac{k}{n}\right) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k} \quad \forall x$$

Demostración.

Definamos $Z_i = I_{(-\infty, X_i]}(x)$. Entonces podemos ver que $Z_i \sim \mathbf{Bernoulli}(F(x))$. Como Z_i depende de una muestra independiente, entonces las Z_i son independientes. De este modo $\sum_{i=1}^n Z_i \sim \mathbf{Bin}(n, F(x))$.

Por lo tanto

$$P\left(\sum_{i=1}^n Z_i = k\right) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

Dividiendo ambos términos de la probabilidad por n obtenemos el resultado. □

Derivado del teorema anterior, se tiene que para una x fija,

$$E[F_n(x)] = F(x)$$

$$Var(F_n(x)) = \frac{F(x)(1 - F(x))}{n}$$

Y por el Teorema del Límite Central:

$$F_n(x) \overset{a}{\sim} \mathcal{N}\left(F(x), \frac{F(x)[1 - F(x)]}{n}\right)$$

$$\therefore \sqrt{n}(F_n(x) - F(x)) \overset{a}{\sim} \mathcal{N}(0, F(x)(1 - F(x)))$$

Para estimar $F(x) \quad \forall x$, podemos utilizar el teorema de Glivenko-Cantelli, que establece la convergencia uniforme de F_n a F :

Teorema (Glivenko-Cantelli)

con probabilidad 1, la convergencia de $F_n(x)$ a $F(x)$ es uniforme:

$$P(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{n \rightarrow \infty} 0) = 1$$

- Consideremos la distancia máxima entre dos distribuciones $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$. Entonces el teorema anterior establece que $P(\lim_{n \rightarrow \infty} D_n = 0) = 1$, así que la distribución de D_n converge a la función que concentra toda su masa en 0.
- Sin embargo, la función de distribución asintótica de $\sqrt{n}D_n$ converge a otra distribución, que ¡no depende de la función de la que la muestra fue obtenida! La distribución límite se conoce como la *distribución de Kolmogorov*², y no la estudiaremos en este curso.
- Tanto D_n como $F_n(X)$ son ejemplos de estadísticas *libre de distribución o no paramétricas*.

²la función de distribución es de la forma: $P(D_n \leq x) = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2 \pi^2 / (8x^2)}$

Pruebas de Uniformidad: Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov (KS) es una *prueba de bondad de ajuste* para distribuciones *continuas*. Formalmente, queremos probar la hipótesis:

$$H_0 : F(x) = F_0(x) \quad \forall x \quad \text{vs.} \quad H_1 : F(x) \neq F_0(x) \quad \text{para alguna } x$$

donde F_0 es la función de distribución objetivo que se supone siguen los datos. La estadística de prueba se basa en la función de distribución empírica.

La estadística de prueba se define como $D_n = \max_x |F_n(x) - F_0(x)|$. Se rechaza la hipótesis nula si D_n es "muy grande".

Una aproximación para muestras grandes ($n \geq 35$) para el p -value es

$$P(D_n > c) \approx 2e^{-2nc^2}$$

donde c se reemplaza por el valor de la estadística obtenida.

Apliquen la prueba de KS a los siguientes 100 números y verifiquen uniformidad:

```
set.seed(1)
uniformes <- runif(100)
uniformes
```

```
[1] 0.26550866 0.37212390 0.57285336 0.90820779 0.20168193 0.89838968 0.94467527 0.66079779 0.62911404 0.06178627
[11] 0.20597457 0.17655675 0.68702285 0.38410372 0.76984142 0.49769924 0.71761851 0.99190609 0.38003518 0.77744522
[21] 0.93470523 0.21214252 0.65167377 0.12555510 0.26722067 0.38611409 0.01339033 0.38238796 0.86969085 0.34034900
[31] 0.48208012 0.59956583 0.49354131 0.18621760 0.82737332 0.66846674 0.79423986 0.10794363 0.72371095 0.41127443
[41] 0.82094629 0.64706019 0.78293276 0.55303631 0.52971958 0.78935623 0.02333120 0.47723007 0.73231374 0.69273156
[51] 0.47761962 0.86120948 0.43809711 0.24479728 0.07067905 0.09946616 0.31627171 0.51863426 0.66200508 0.40683019
[61] 0.91287592 0.29360337 0.45906573 0.33239467 0.65087047 0.25801678 0.47854525 0.76631067 0.08424691 0.87532133
[71] 0.33907294 0.83944035 0.34668349 0.33377493 0.47635125 0.89219834 0.86433947 0.38998954 0.77732070 0.96061800
[81] 0.43465948 0.71251468 0.39999437 0.32535215 0.75708715 0.20269226 0.71112122 0.12169192 0.24548851 0.14330438
[91] 0.23962942 0.05893438 0.64228826 0.87626921 0.77891468 0.79730883 0.45527445 0.41008408 0.81087024 0.60493329
```

KS: Pasos a seguir (caso uniforme) I

1. Calcular las estadísticas de orden de la muestra $r_{(i)}$. Sea n el número de valores.

```
unif_ord <- sort(uniformes,decreasing=F)
head(unif_ord,20)

[1] 0.01339033 0.02333120 0.05893438 0.06178627 0.07067905 0.08424691 0.09946616 0.10794363 0.12169192 0.12555510
[11] 0.14330438 0.17655675 0.18621760 0.20168193 0.20269226 0.20597457 0.21214252 0.23962942 0.24479728 0.24548851
```

2. Calculen

$$D^+ = \max_i \left\{ \frac{i}{n} - r_{(i)} \right\}$$

$$D^- = \max_i \left\{ r_{(i)} - \frac{i-1}{n} \right\}$$

(Valores correspondientes a $F_0 \equiv \mathcal{U}(0, 1)$).

```
Dp <- max(1:length(unif_ord)/length(unif_ord)-unif_ord)
Dp

[1] 0.04712408

Dm <- max(unif_ord - 0:(length(unif_ord)-1)/length(unif_ord))
Dm

[1] 0.07627171
```

KS: Pasos a seguir (caso uniforme) II

3. Calculen $D = \max(D^+, D^-)$

```
Dmax <- max(Dp,Dm)
Dmax
[1] 0.07627171
```

4. Usen una tabla apropiada para encontrar el valor crítico D_α de la distribución de D . Rechazas H_0 si $D > D_\alpha$ (o mejor aún, obtén un p -value).

```
pval <- 2*exp(-2*length(unif_ord)*Dmax^2)
pval
[1] 0.6247976
```

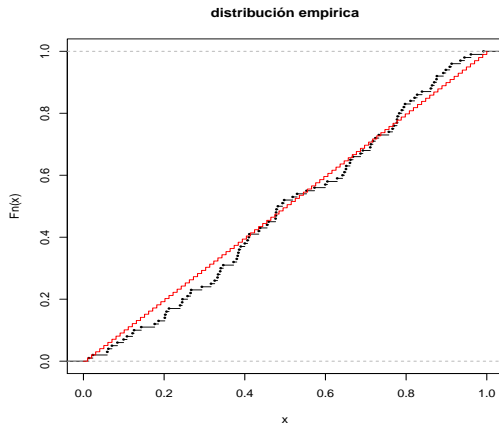
En R se pueden usar las funciones `ecdf` para obtener la función de distribución empírica, y `ks.test` para hacer la prueba de Kolmogorov-Smirnov.

```
# antes vemos la distribución empírica  
Fn <- ecdf(unif_ord)  
summary(Fn) #resumen de los puntos generados.
```

```
Empirical CDF: 100 unique values with summary  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
0.01339 0.32308 0.48781 0.51785 0.76719 0.99191
```

KS en R

```
plot(Fn, main = "distribución empírica", xlim = c(0,1), ylim = c(0,1), pch = 16, cex = 0.5)  
sq <- seq(0, 1, length = 100)  
lines(sq, punif(sq), type = "s", col = "red", lwd = 1)
```



En un sólo paso:

```
ks.test(unif_ord,"punif")
```

One-sample Kolmogorov-Smirnov test

data: unif_ord

D = 0.076272, p-value = 0.6058

alternative hypothesis: two-sided

Solo para corroborar el calculo de D:

```
sq <- seq(0,1,length=100000)
```

```
max(abs(Fn(sq)-punif(sq)))
```

```
[1] 0.07626316
```

Hay otras pruebas similares a la prueba KS. Todas ellas comparan la función de distribución empírica con la teórica utilizando diferentes métricas funcionales.

- Kolmogorov-Smirnov: D_n
- Cramér-von Misses:

$$W_n = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x)$$

- Anderson-Darling

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2 dF(x)}{F(x)(1 - F(x))}$$

Sin embargo, de acuerdo a [L'Ecuyer & Simard \(2007\)](#), estas últimas dos estadísticas no son tan potentes como la prueba de KS para probar aleatoriedad.

Pruebas alternativas a KS

Ejemplo:

```
library(goftest)
cvm.test(uniformes) #Cramer-von Mises

Cramer-von Mises test of goodness-of-fit
Null hypothesis: uniform distribution
Parameters assumed to be fixed

data: uniformes
omega2 = 0.12243, p-value = 0.4864

ad.test(uniformes) #Anderson-Darling

Anderson-Darling test of goodness-of-fit
Null hypothesis: uniform distribution
Parameters assumed to be fixed

data: uniformes
An = 0.85484, p-value = 0.4428
```


Pruebas de uniformidad: Prueba de χ^2 de Pearson

La prueba de χ^2 de Pearson (1900) es la primera prueba de bondad de ajuste; incluso es una de las primeras pruebas de inferencia estadística.

- La hipótesis estadística a probar es la misma que la de la prueba de KS:

$$H_0 : F(x) = F_0(x) \quad \forall x \text{ vs. } H_a : F(x) \neq F_0(x) \text{ para alguna } x$$

- La prueba “compara” el histograma obtenido de los datos observados con la verdadera densidad de la distribución supuesta de los datos.
- La prueba es mucho más conveniente para distribuciones discretas que para distribuciones continuas.
- Versión ‘discreta’ de la prueba de K-S.

Procedimiento para χ^2

- 1 Particiona el rango de la distribución supuesta en k subintervalos con límites $\{a_0, a_1, \dots, a_k\}$, y define a N_j como el número de observaciones en cada subintervalo, para cada j .
- 2 Calcular la proporción esperada de observaciones en el intervalo $(a_{j-1}, a_j]$ como $p_j = \int_{a_{j-1}}^{a_j} dF(x)$.
- 3 La estadística de prueba es

$$\chi^2 = \sum_{i=1}^k \frac{(N_j - np_j)^2}{np_j}.$$

Se rechaza la hipótesis nula si χ^2 es grande, considerando que $\chi^2 \overset{a}{\sim} \chi_{k-1}^2$.

χ^2 Ejemplo

La siguiente función en R hace la prueba descrita.

```
prueba.chisq.uniforme <- function(x, k = ceiling(length(x)/5)){  
  n <- length(x)  
  part <- seq(0, 1, length = k + 1) #partición  
  z <- hist(x, breaks = part, plot = F)$counts  
  ch <- (k/n)*sum((z-n/k)^2) #estadística chi  
  pval <- pchisq(ch, k - 1, lower.tail = F)  
  return(list(part = part, freqs = z, estadística = ch, pval = pval))  
}
```

Apliquen la prueba a los mismos datos del ejemplo anterior. ¿Qué se concluye?

```
prueba.chisq.uniforme(uniformes, k = 20)  
  
$part  
[1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00  
  
$freqs  
[1] 2 5 4 2 7 4 7 7 5 9 2 3 4 7 5 10 4 7 4 2  
  
$estadística  
[1] 21.2  
  
$pval  
[1] 0.3258558
```

La función `chisq.test` en R hace la prueba descrita, pero requerimos pasarle como parámetro las probabilidades de la distribución objetivo:

```
h1 <- hist(uniformes, breaks = 20, right = F, plot = F)
#crea las probabilidades esperadas sobre la partición creada:
breaks_cdf <- punif(h1$breaks)
null.probs <- breaks_cdf[-1] - breaks_cdf[-length(breaks_cdf)]
a <- chisq.test(h1$counts, p = null.probs, rescale.p = T,
               simulate.p.value = T)
a
```

Chi-squared test for given probabilities with simulated p-value (based on 2000 replicates)

data: h1\$counts
X-squared = 21.2, df = NA, p-value = 0.3263

Observaciones sobre KS y χ^2

- Dado que la prueba es asintótica, se requiere un tamaño grande de muestra ($n \geq 100$ y $n/k \geq 5$ por celda en principio).
- La prueba de KS es más potente que la prueba de χ^2 y puede ser aplicada a muestras más pequeñas.

Gráficas de probabilidad o qq -plots

Las gráficas de probabilidad o qq -plots (quantile-quantile plots) comparan los cuantiles de la muestra contra los cuantiles teóricos de la población.

- Un p -cuantíl o p -percentíl es un número x_p tal que $F(x_p) = P(X < x_p) = p$. Para distribuciones discretas, no necesariamente es único, y usualmente se redondea al entero más cercano.
- Una gráfica cuantíl-cuantíl o qq -plot consiste de los puntos $(X_{(i)}, q_i)$, donde q_i es el $\frac{i}{n}$ -cuantíl de la distribución objetivo.

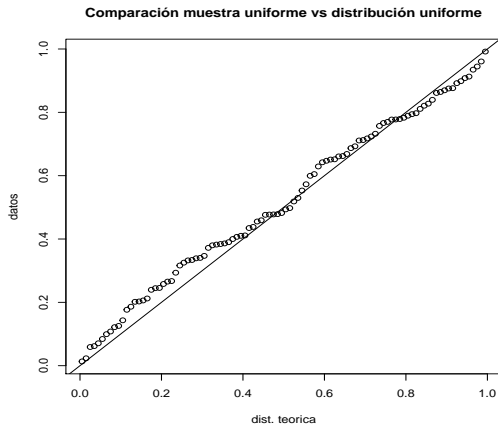
Se recomienda utilizar el cuantíl $\frac{i-0.5}{n}$ en lugar de $\frac{i}{n}$ como corrección por continuidad.

qq—plot Ejemplo I

- Si el qq—plot sigue la recta identidad cuando se grafica contra la distribución teórica, entonces se puede decir que los datos siguen adecuadamente la distribución objetivo.
- Sin embargo, esta no es una prueba estadística, sólo una guía visual.
- La siguiente función crea gráficas de probabilidad para cualquier distribución:

```
graf.teorica <- function(fun.quan, x, tit, ...){  
  z <- sort(x, decreasing = F)  
  plot(fun.quan(ppoints(z),...), z, main = tit,  
        xlab = "dist. teorica", ylab = "datos")  
  abline(a = 0, b = 1)}  
graf.teorica(qunif, uniformes, tit = "Comparación muestra uniforme vs distribución uniforme")
```

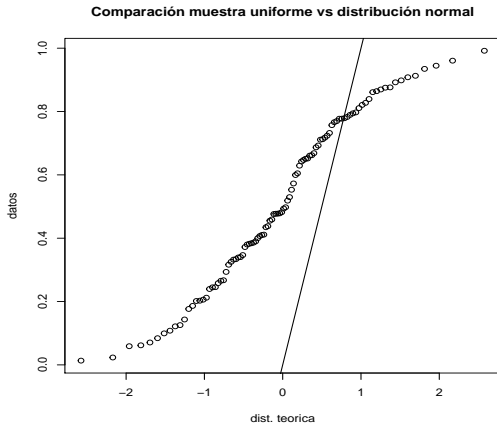
qq-plot Ejemplo II



qq-plot Ejemplo 2 I

- Comparamos ahora en relación a la distribución normal estándar para ver la desviación:

```
graf.teorica(qnorm, uniformes, tit = "Comparación muestra uniforme vs distribución normal",0,1)
```



Pruebas de independencia

Las pruebas de independencia que revisaremos son las siguientes:

- Rachas (signos, runs, etc.)
- Prueba de gaps
- Prueba de poker
- Autocorrelación

La mayoría de estas pruebas son no paramétricas o libres de distribución.

Racha

Una *racha* es una sucesión continua de eventos similares de un tipo (tipo 1) y seguida por sucesiones de otros eventos de otro tipo (tipo 2). La longitud de la racha es el número de eventos similares en esa sucesión. En esta definición se consideran sólo dos tipos de eventos, pero puede haber más de dos tipos en general.

- Un ejemplo de racha es el siguiente, con eventos A y B : AA BBB AA B AAA BBBB A BB. En este ejemplo, se tienen 8 rachas (se cuentan todas).
- Ejemplos de rachas: números crecientes o números decrecientes; en volados: águila o sol, autos amarillos que pasan por un crucero vs otro tipo de autos, etc.

La teoría general y la prueba para rachas de más de dos tipos se puede encontrar en el siguiente [paper](#): Mood, A. M., *The distribution theory of runs* Ann. Math. Statist. Volume 11, Number 4 (1940), 367-392.

Prueba de rachas I

- Las rachas tratan de identificar patrones en el acomodo sucesivo de las observaciones. El siguiente patrón intuitivamente no es aleatorio:

ABABABABABAB...

Aquí podemos considerar que tenemos un total $R = 12$ rachas, con $R_A = 6$ rachas para A y $R_B = 6$ para B . Otro patrón no aleatorio es el siguiente, en donde se forman dos conglomerados de tipos de eventos:

AAAAAABBBBBB...

- Las rachas pueden analizarse desde dos puntos de vista, dando origen a diferentes pruebas (lo que a veces añade confusión):
 - por el número de rachas y
 - por la longitud de las rachas de cada tipo.
 - por combinaciones de los dos casos anteriores.

- Cada uno de estos o la combinación de ellos se puede usar para probar la hipótesis de aleatoriedad/independencia. Las siguientes podrían ser criterios para rechazar independencia:
 - muy pocas rachas
 - demasiadas rachas
 - demasiadas rachas de longitud grande, etc.
- Los datos se pueden dicotomizar artificialmente para formar los dos grupos que se consideran para la formación de rachas:
 - comparación respecto a un valor focal (media, mediana, etc.)
 - rachas crecientes o decrecientes

Prueba de rachas: Ejemplo

- Para el siguiente conjunto de 10 datos:

0.86, 0.11, 0.23, 0.03, 0.13, 0.06, 0.55, 0.64, 0.87, 0.10

Si consideramos su mapeo a **rachas crecientes**, tomando los signos de las diferencias $x_{n+1} - x_n$ como los dos objetos que forman las rachas:

$$0.11 - 0.86 = -; 0.23 - 0.11 = +; 0.03 - 0.23 = -; 0.13 - 0.03 = +; 0.06 - 0.13 = -;$$

$$0.55 - 0.06 = +; 0.64 - 0.55 = +; 0.87 - 0.64 = +; 0.10 - 0.87 = -$$

Se obtiene la nueva serie: $- + - + - + + + -$, que tiene 9 elementos (uno menos que el número de datos), y hay 7 rachas, con seis rachas de longitud 1 y una de longitud 3.

- Noten que con n datos, el máximo número de *rachas crecientes* es $n - 1$ y el mínimo es 1.

Modelo para prueba de rachas basada en su número

- En un caso más general que el de rachas crecientes o decrecientes, la prueba de rachas puede ser usada como una prueba de aleatoriedad.
- Consideren una sucesión de n elementos, con n_1 elementos del tipo 1 y n_2 elementos del tipo 2. Entonces $n = n_1 + n_2$.
- Sea R_i el número de rachas de tipo i para $i = 1, 2$, y $R = R_1 + R_2$ el total de rachas. Estas son consideradas variables aleatorias. ¿Cuál es la distribución de cada R_i y del total R ? Para obtener estas distribuciones, tenemos que repasar algunos resultados de combinatoria.
- Con estas distribuciones, se puede calcular una prueba para la hipótesis

H_0 : la muestra es aleatoria vs H_a : la muestra no es aleatoria

- Bajo H_0 cada arreglo de los $n_1 + n_2$ elementos tiene la misma probabilidad. Entonces el número total de arreglos distinguibles es
$$\binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2} = (n_1 + n_2)! / n_1! n_2!.$$
- Para calcular la probabilidad de observar $R_1 = r_1$ y $R_2 = r_2$, ya tenemos el denominador, pero para el numerador, usamos el siguiente lema.

Lema previo

Lema 1.

El número de formas distinguibles de acomodar n objetos no distinguibles en r celdas distinguibles sin celdas vacías es $\binom{n-1}{r-1}$, $n \geq r$.

Demostración.

Supongamos que los objetos indistinguibles son asteriscos (*). Se colocan los n asteriscos en una fila y para poner las r celdas, se insertan $r - 1$ divisiones entre cualesquiera dos asteriscos en la línea.

Eg: si $n = 5$ y $r = 4$: $*|**|*|*$. Aquí hay 4 posibles lugares entre los asteriscos, en donde se pueden insertar 3 divisiones para simular 4 celdas.

En esta configuración, habrá $n - 1$ posiciones en las que las $r - 1$ divisiones pueden insertarse. Entonces el problema es equivalente al número de subconjuntos de tamaño $r - 1$ de un total de $n - 1$ elementos, es decir, $\binom{n-1}{r-1}$ posibles acomodados. □

Distribución conjunta de R_1 y R_2 I

- Entonces, para calcular la distribución conjunta de R_1 y R_2 dada por $P(R_1 = r_1, R_2 = r_2)$, el numerador es el número de arreglos distinguibles de n objetos con r_1 rachas tipo 1 y r_2 rachas tipo 2.
- Los números r_1 y r_2 están “amarrados”, ya que las rachas dependen de los cambios de elementos. Sólo podemos tener que $r_1 = r_2$ o bien, que $r_1 = r_2 \pm 1$, dependiendo de con qué elemento empezamos a contar las rachas.

Distribución conjunta de R_1 y R_2

La distribución conjunta de R_1 y R_2 es

$$f_{R_1, R_2}(r_1, r_2) = \frac{c \binom{n_1-1}{r_1-1} \binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n_1}} \quad r_i \in \{1, 2, \dots, n_i\}, r_1 = r_2 \text{ ó } r_1 = r_2 \pm 1$$

donde $c = 2$ si $r_1 = r_2$, o $c = 1$ si $r_1 = r_2 \pm 1$.

Distribución conjunta de R_1 y R_2 II

Demostración.

- Notar que bajo H_0 , que supone aleatoriedad, cada posible arreglo de los n elementos es **equiprobable**, con probabilidad $1/\binom{n_1+n_2}{n_1} = 1/\binom{n_1+n_2}{n_2}$.
- Para obtener una sucesión con r_1 rachas de objetos tipo 1, los n_1 objetos deben ser colocados en r_1 celdas. Por el resultado anterior, esto se puede hacer de $\binom{n_1-1}{r_1-1}$ formas. El mismo argumento aplica si se requieren r_2 rachas de objetos tipo 2.
- Ya empaquetados los datos en rachas, el número de arreglos distinguibles comenzando con una racha de tipo 1 es $\binom{n_1-1}{r_1-1} \binom{n_2-1}{r_2-1}$. Similarmente para una sucesión que empieza con una racha tipo 2.
- Como las rachas alternan por definición, necesariamente o $r_1 = r_2 \pm 1$ o bien, $r_1 = r_2$. Si $r_1 = r_2 + 1$, la sucesión debe comenzar con una racha tipo 1. Si $r_1 = r_2 - 1$, la sucesión empieza con una racha tipo 2. Si $r_1 = r_2$ se puede comenzar con cualquier racha, por lo que el número de arreglos se duplica.



Distribución marginal de R_i I

Distribución marginal de R_1

$$f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1} \binom{n_2+1}{r_1}}{\binom{n_1+n_2}{n_1}} \quad r_1 \in \{1, 2, \dots, n_1\}$$

Similarmente, se obtiene la distribución marginal para R_2 , intercambiando en la ecuación anterior los valores de n_1 y n_2 .

Demostración.

Como $r_2 \in \{r_1 - 1, r_1, r_1 + 1\}$, tenemos que $f_{R_1}(r_1) = \sum_{r_2} f_{R_1, R_2}(r_1, r_2)$. Así que:

$$\begin{aligned}\binom{n_1 + n_2}{n_1} f_{R_1}(r_1) &= 2 \binom{n_1 - 1}{r_1 - 1} \binom{n_2 - 1}{r_1 - 1} + \binom{n_1 - 1}{r_1 - 1} \binom{n_2 - 1}{r_1 - 2} + \binom{n_1 - 1}{r_1 - 1} \binom{n_2 - 1}{r_1} \\&= \binom{n_1 - 1}{r_1 - 1} \left[\binom{n_2 - 1}{r_1 - 1} + \binom{n_2 - 1}{r_1 - 2} + \binom{n_2 - 1}{r_1 - 1} + \binom{n_2 - 1}{r_1} \right] \\&= \binom{n_1 - 1}{r_1 - 1} \left[\binom{n_2}{r_1 - 1} + \binom{n_2}{r_1} \right] \\&= \binom{n_1 - 1}{r_1 - 1} \binom{n_2 + 1}{r_1}\end{aligned}$$



Distribución de $R = R_1 + R_2$ I

Finalmente, juntando los resultados anteriores, podemos obtener la distribución exacta del número total de rachas R :

Distribución del número total de rachas R

$$P(R = r) = \begin{cases} \frac{2 \binom{n_1-1}{k-1} \binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}, & r = 2k \\ \frac{\binom{n_1-1}{k} \binom{n_2-1}{k-1} + \binom{n_2-1}{k} \binom{n_1-1}{k-1}}{\binom{n_1+n_2}{n_1}}, & r = 2k + 1 \end{cases}$$

Distribución de $R = R_1 + R_2$ II

Demostración.

Si el número de rachas r es par, entonces es de la forma $r = 2k$ para algún número natural k y se tiene el mismo número de rachas de ambos tipos. Los posibles valores de r_1 y r_2 son entonces $r/2 = r_1 = r_2$ y la distribución conjunta se suma sobre este par de valores. Haciendo $k = r/2$ se obtiene el resultado.

Si r impar, entonces $r = 2k \pm 1$. En este caso, la densidad conjunta se suma sobre los dos pares de valores:

- $r_1 = \frac{r-1}{2} = k$ y $r_2 = \frac{r+1}{2} = k + 1$, y
- $r_1 = \frac{r+1}{2} = k + 1$ y $r_2 = \frac{r-1}{2} = k$. Haciendo las sustituciones correspondientes se obtiene el resultado.



Distribución de $R = R_1 + R_2$ III

Un poco (bueno, no tanto) más elaborado de probar, es el hecho de que:

$$\begin{aligned}E(R) &= 1 + \frac{n_1 n_2}{n_1 + n_2} \\ \text{Var}(R) &= \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}\end{aligned}$$

Ejemplo: prueba de hipótesis de independencia. I

Con los resultados anteriores, podemos probar la hipótesis de aleatoriedad. la función `fr` en el código de R es la distribución del número total de rachas que acabamos de obtener. Esta función también está en `randtests` como funciones `*Runs`.

Si $n_1 = 5$ y $n_2 = 4$, tenemos entonces que:

```
fr <- function(n1,n2,r){  
  if(r<2 || r>n1+n2) error("r tiene que ser mayor a 1 o menor que la suma de n1 y n2")  
  if( r %% 2 ==0) {  
    sol <- 2*choose(n1-1,r/2-1)*choose(n2-1,r/2-1)/choose(n1+n2,n1)  
  } else {  
    sol <- (choose(n1-1,(r-1)/2)*choose(n2-1,(r-3)/2) +  
           choose(n1-1,(r-3)/2)*choose(n2-1,(r-1)/2))/choose(n1+n2,n1)  
  }  
  return(sol)  
}
```

Ejemplo: prueba de hipótesis de independencia. II

```
fr(n1=5,n2=4,r=9) #equivalente a druns(9,5,4)
```

```
[1] 0.007936508
```

```
fr(n1=5,n2=4,8)
```

```
[1] 0.06349206
```

```
fr(n1=5,n2=4,2)
```

```
[1] 0.01587302
```

```
fr(n1=5,n2=4,3)
```

```
[1] 0.05555556
```

Para una prueba de dos lados que rechace la hipótesis nula para $R \leq 2$ o $R \geq 9$, el nivel de significancia exacto sería $f_R(2) + f_R(9) = 0.024$

Para la región crítica definida por $R \leq 3$ o $R \geq 8$, sería $\alpha = 18/126 = 0.143$

Prueba de rachas para muestras grandes

- La distribución de R se puede conocer de manera exacta, pero comienza a ser difícil de manejar si n es grande. Se puede usar una aproximación asintótica bajo ciertos supuestos cuando $n_i \geq 10$.
- Supongamos que $n_1 \rightarrow \infty$ de tal forma que $\frac{n_1}{n} \rightarrow \lambda$ y $\frac{n_2}{n} \rightarrow 1 - \lambda$. Entonces

$$\begin{aligned} E(R/n) &\rightarrow 2\lambda(1 - \lambda), \text{ y} \\ \text{Var}(R/\sqrt{n}) &\rightarrow 4\lambda^2(1 - \lambda)^2 \end{aligned}$$

Así que

$$Z = \frac{R - 2n\lambda(1 - \lambda)}{2\sqrt{n}\lambda(1 - \lambda)}$$

es asintóticamente $\mathcal{N}(0, 1)$.

Ejemplo

Considerando la siguiente muestra de números pseudoaleatorios:

```
set.seed(100)
x <- runif(50); head(x)

[1] 0.30776611 0.25767250 0.55232243 0.05638315 0.46854928 0.48377074

#función para contar las rachas:
nrachas <- function(x){
  n <- length(x)
  signo <- x[-1] - x[-n]
  n1 <- sum(signo < 0)
  s <- ifelse(signo < 0, -1, 1)
  R <- 1 + sum(s[-1] != s[-(n-1)]) #cuenta los cambios de signo
  return(list(R=R, n1=n1, lambda=n1/length(x)))
}
res <- nrachas(x); res

$R
[1] 34

$n1
[1] 21

$lambda
[1] 0.42

lambda <- res$lambda
z <- (nrachas(x)$R - 2*length(x)*lambda*(1-lambda))/(2*sqrt(length(x)) * lambda * (1-lambda));z

[1] 2.798239

pnorm(z)

[1] 0.9974309
```

Entonces $z_0 = 2.7982387$, y el p-value es 0.9974309.

Rachas arriba o abajo de un umbral λ

La prueba de rachas basada en el número de rachas puede no estar observando todas las posibles condiciones para aleatoriedad cuando se considera el número de rachas. Por ejemplo, la siguiente sucesión pasa la prueba:

```
x <- c(0.63, 0.72, 0.79, 0.81, 0.52, 0.94, 0.83, 0.93, 0.87, 0.67,  
      0.54, 0.83, 0.89, 0.55, 0.88, 0.77, 0.74, 0.95, 0.82, 0.86,  
      0.43, 0.32, 0.36, 0.18, 0.08, 0.19, 0.18, 0.27, 0.36, 0.34,  
      0.31, 0.45, 0.49, 0.43, 0.46, 0.35, 0.25, 0.39, 0.47, 0.41)  
z <- ( nrachas(x)$R - 2*length(x)*lambda*(1-lambda) )/( 2*sqrt(length(x))*lambda*(1-lambda) )  
pnorm(z)  
  
[1] 0.9827158
```

Sin embargo, los primeros 20 números están por encima de la media 0.5565 y los otros 20 por debajo, lo cual es altamente improbable si los números son aleatorios. La prueba de rachas se puede adaptar para considerar posibles tendencias en su comportamiento, considerando la siguiente modificación.

Rachas arriba o abajo de un umbral II

Para n “grande” ($n > 20$), la distribución asintótica de R tiene media y varianza dada por:

$$\begin{aligned}\mu_R &= \frac{2n_1n_2}{n} + \frac{1}{2} \\ \sigma_R^2 &= \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}\end{aligned}$$

donde n es el número de datos en la muestra, n_1 es el número de observaciones sobre la media, y n_2 el número de observaciones por debajo de la media. Si $n > 20$, la distribución de R es aproximadamente $\mathcal{N}(\mu_R, \sigma_R^2)$. La estadística de prueba es

$$z_0 = \frac{R - (2n_1n_2)/n - 1/2}{\sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}} \sim \mathcal{N}(0, 1)$$

Distribución asintótica de la prueba de rachas crecientes o decrecientes I

La distribución asintótica para la prueba de rachas basada en rachas crecientes es de la forma:

Para n “grande” ($n > 20$), la distribución asintótica de R tiene media y varianza dada por:

$$\begin{aligned}\mu_R &= \frac{2n-1}{3} \\ \sigma_R^2 &= \frac{16n-29}{90}\end{aligned}$$

donde n es el número de datos en la muestra.

Si $n > 20$, la distribución de R es aproximadamente $\mathcal{N}(\mu_R, \sigma_R^2)$. La estadística de prueba es

$$z_0 = \frac{R - (2n-1)/3}{\sqrt{\frac{16n-29}{90}}} \sim \mathcal{N}(0, 1)$$

Prueba de Rachas de Levene y Wolfowitz (1944) I

Basada en la longitud de las rachas

Una versión más de la prueba de rachas considera el caso de rachas crecientes o decrecientes, junto con la longitud de las diferentes rachas. Por ejemplo:

0.134, 0.279, 0.866, 0.197, 0.011, 0.923, 0.990, 0.876

- Sea R_k el número de rachas de longitud k . El total de rachas es $R = R_1 + R_2 + \dots$
- Levene y Wolfowitz mostraron en este caso que en una sucesión de n $U(0, 1)$, el número esperado de rachas crecientes de longitud $k \geq 1$ está dada por:

$$E(R_k) = \frac{(k^2 + k - 1)(n - k - 1)}{(k + 2)!}, \quad 1 \leq k \leq n$$

- Como usualmente n es grande, se puede considerar la aproximación:

$$E(R_k) \approx \frac{(k^2 + k - 1)}{(k + 2)!} n, \quad k \ll n$$

Prueba de Rachas de Levene y Wolfowitz (1944) II

Basada en la longitud de las rachas

- Entonces, para n fija $E(R_k)$ decrece conforme $k \rightarrow n$ y es usual considerar la distribución conjunta de $(R_1, R_2, \dots, R_j, S_j)$ para alguna $j > 1$, donde $S_j = \sum_{k=j+1}^n R_k$. Usualmente se considera $j = 5$.
- Una observación importante aquí es que las R'_i s no son independientes, y por lo tanto no se puede aplicar la prueba χ^2 de directa.
- Si n es el número de observaciones, y se define a r_6 como el número de rachas que son de longitud igual o mayor a 6, entonces la estadística de Levene-Wolfowitz es:

$$R_n = \frac{1}{n} \sum_{i=1}^6 \sum_{j=1}^6 a_{ij} (r_i - nb_i)(r_j - nb_j) = \frac{1}{n} (\mathbf{r} - \mathbf{nb})' \mathbf{a} (\mathbf{r} - \mathbf{nb})$$

donde $\mathbf{a}_{6 \times 6}$ es una matriz simétrica y $\mathbf{b}_{6 \times 1}$ son constantes, que provienen de las estadísticas de orden. La matriz está dada por

Prueba de Rachas de Levene y Wolfowitz (1944) III

Basada en la longitud de las rachas

$$\mathbf{a} = \begin{pmatrix} 4,529.4 & 9,044.9 & 13,568 & 18,091 & 22,615 & 27,892 \\ & 18,097 & 27,139 & 36,187 & 45,234 & 55,789 \\ & & 40,721 & 54,281 & 67,852 & 83,685 \\ & & & 72,414 & 90,470 & 111,580 \\ & & & & 113,262 & 139,476 \\ & & & & & 172,860 \end{pmatrix}$$

$$\mathbf{y} \mathbf{b} = \left(\frac{1}{6}, \frac{5}{24}, \frac{11}{120}, \frac{19}{720}, \frac{29}{5040}, \frac{1}{840} \right).$$

- Se puede probar que la estadística tiene una distribución asintótica χ^2_6 . (D. Knuth: *The Art of Computer Programming, Vol.2*) Se recomienda $n \geq 4,000$.

Prueba de rachas: Ejemplo I

La función siguiente devuelve el valor de la estadística y los valores de r_i , así como el p -value de la prueba. Por ejemplo, si se genera una muestra aleatoria:

```
set.seed(1)
x <- runif(5000) #genera una muestra de 5000 números aleatorios
prueba.rachas <- function(x){
a <- matrix(c(4529.4, 9044.9, 13568, 18091, 22615, 27892,
             9044.9, 18097, 27139, 36187, 45234, 55789,
             13568, 27139, 40721, 54281, 67852, 83685,
             18091, 36187, 54281, 72414, 90470, 111580,
             22615, 45234, 67852, 90470, 113262, 139476,
             27892, 55789, 83685, 111580, 139476, 172860), nrow = 6)

b <- c(1/6,5/24,11/120,19/720,29/5040,1/840)
n <- length(x)
x1 <- c(1) #inicializa el indicador de cambios de signo
x1[2:length(x)] <- sign(diff(x)) #guardamos los cambios de signos de la muestra
cambios <- c((1:length(x1))[x1==1],length(x)+1) #contamos los cambios de signo
tabla <- table(c(cambios[1]-1,diff(cambios)),exclude=NULL)
aa <- tabla[match(1:length(x),as.numeric(names(tabla)))]
aa <- ifelse(is.na(aa),0,aa)
aa[6] <- sum(aa[6:n]) #agrupa el número de rachas de longitud 6 o más
r <- aa[1:6]
names(r) <- c(1:5,">=6")
R <- as.numeric((r-n*b)%*%a)%*t(t((r-n*b)))/n) #Estadística de Levene-Wolfowitz
return(list(x=head(x,10),R=R,r=r,pval=pchisq(R,6,lower.tail=F)))
}
```

Prueba de rachas: Ejemplo II

```
prueba.rachas(x)
```

```
$x
```

```
[1] 0.26550866 0.37212390 0.57285336 0.90820779 0.20168193 0.89838968 0.94467527 0.66079779 0.62911404 0.06178627
```

```
$R
```

```
[1] 6.999105
```

```
$r
```

```
  1   2   3   4   5 >=6  
831 996 502 125  28   5
```

```
$pval
```

```
[1] 0.32093
```

Prueba de rachas: Ejemplo II

En el paquete `randtests` hay una versión de la prueba de rachas (Wald-Wolfowitz, altas y bajas con respecto a un umbral). También hay una versión de la prueba en el paquete `snpar` (supplementary Non-parametric statistics methods).

```
library(randtests)
runs.test(x, threshold = mean(x))
```

Runs Test

```
data: x
statistic = -0.14035, runs = 2495, n1 = 2449, n2 = 2551, n = 5000, p-value = 0.8884
alternative hypothesis: nonrandomness
```

```
runs.test(x) #por default es la mediana
```

Runs Test

```
data: x
statistic = 0, runs = 2501, n1 = 2500, n2 = 2500, n = 5000, p-value = 1
alternative hypothesis: nonrandomness
```

Prueba de gaps

- La prueba de gaps (o “huecos”) investiga la relevancia del intervalo entre la recurrencia de un mismo dígito. Lo interesante es medir la longitud L de los gaps para un cierto dígito.
- Por ejemplo, en la siguiente serie, la longitud de los gaps asociados con el 6 se puede determinar:

```
x <- c(
1, 3, 7, 4, 8, 6, 2, 5, 1, 6, 4, 4, 3, 3, 4, 2, 1, 5, 8, 7,
0, 7, 6, 2, 6, 0, 5, 7, 8, 0, 1, 1, 2, 6, 7, 6, 3, 7, 5, 9,
0, 8, 8, 2, 6, 7, 8, 1, 3, 5, 3, 8, 4, 0, 9, 0, 3, 0, 9, 2,
2, 3, 6, 5, 6, 0, 0, 1, 3, 4, 4, 6, 9, 9, 8, 5, 6, 0, 1, 7,
5, 6, 7, 9, 4, 9, 3, 1, 8, 3, 3, 6, 6, 7, 8, 2, 3, 5, 9, 6,
6, 7, 0, 3, 1, 0, 2, 4, 2, 0, 6, 4, 0, 3, 9, 3, 6, 8, 1, 5)
table(x)

x
 0  1  2  3  4  5  6  7  8  9
14 11 10 16 10 10 18 11 11  9
```

- En este ejemplo, hay 18 números ‘6’ que se repiten, y los gaps son los siguientes: el primero es de longitud 3, el segundo es de longitud 12, el tercero es de longitud 1, etc.

Prueba de gaps I

- En general, la probabilidad de un gap de longitud x para el dígito U está dada por:

$$P(L = x) = P(U \text{ seguido de exactamente } x \text{ dígitos no } U) = (0.1)(0.9)^x$$

para $x = 0, 1, 2, \dots$

- Para llevar a cabo la prueba de independencia, se tienen que obtener todos las longitudes de los gaps de todos los dígitos y analizarlos, aplicando alguna prueba de bondad de ajuste como la prueba de KS o la de χ^2 .
- La función de distribución teórica para los dígitos es:

$$P(L \leq x) = F_L(x) = 0.1 \sum_{j=0}^x (0.9)^j = 1 - 0.9^{x+1}$$

Prueba de gaps: ejemplo I

- Con los datos provistos anteriormente, y para el caso del dígito 6, se tienen 17 gaps siguientes:

```
gaps <- function(x){ #calcula todos las longitudes de gaps de la serie x
  lgaps <- NULL
  for (i in 0:9){
    pos <- which(x==i)
    l <- diff(pos)
    lgaps <- c(lgaps,l-1)}
  L <- table(lgaps)
  return(L)
}
l <- gaps(x)
cumsum(l/sum(l))
```

0	1	2	3	4	5	6	7	8	9	10	11
0.1000000	0.1909091	0.2454545	0.3000000	0.3545455	0.4000000	0.4545455	0.5090909	0.5636364	0.6181818	0.6909091	0.7181818
12	13	14	15	16	17	19	21	22	24	33	34
0.7545455	0.8181818	0.8272727	0.8545455	0.8909091	0.9090909	0.9181818	0.9272727	0.9636364	0.9727273	0.9818182	0.9909091
37											
1.0000000											

Prueba de gaps: ejemplo II

- Para calcular las frecuencias teóricas, podemos usar la distribución geométrica.

```
pgeom(as.numeric(names(l)),prob=0.1)
[1] 0.1000000 0.1900000 0.2710000 0.3439000 0.4095100 0.4685590 0.5217031 0.5695328 0.6125795 0.6513216 0.6861894
[12] 0.7175705 0.7458134 0.7712321 0.7941089 0.8146980 0.8332282 0.8499054 0.8784233 0.9015229 0.9113706 0.9282102
[23] 0.9721872 0.9749684 0.9817520

(D <- max(abs(cumsum(1/sum(l))-pgeom(as.numeric(names(l)),prob=0.1))))
[1] 0.068559

(pval <- 2*exp(-2*sum(l)*D^2))
[1] 0.7111109
```

Prueba de gaps a números aleatorios

- Cuando se aplica la prueba de gaps a números aleatorios, se utilizan clases de intervalos para representar a los dígitos.
- Por ejemplo, se pueden considerar una partición del intervalo unitario en los intervalos:

$$[0, 0.1), [0.1, 0.2), \dots, [0.9, 1]$$

- Entonces se asocia el dígito que corresponde a cada intervalo y se utiliza el conjunto de dígitos obtenido.

Prueba de gaps: ejemplo

Realizar la prueba de gaps para la siguiente lista de datos:

```
set.seed(1)  
(x <- runif(10))
```

```
[1] 0.26550866 0.37212390 0.57285336 0.90820779 0.20168193 0.89838968 0.94467527 0.66079779 0.62911404 0.06178627
```

Prueba de gaps ejemplo

- En el paquete randtoolbox se tiene la prueba de gaps

```
library(randtoolbox)
Loading required package: rngWELL
This is randtoolbox. For an overview, type 'help("randtoolbox")'.

Attaching package: 'randtoolbox'
The following object is masked from 'package:randtests':
  permut

x
[1] 0.26550866 0.37212390 0.57285336 0.90820779 0.20168193 0.89838968 0.94467527 0.66079779 0.62911404 0.06178627
gap.test(x)

Gap test

chisq stat = 1.1, df = 3, p-value = 0.77

(sample size : 10)

length observed freq theoretical freq
1  2  1.2
2  1  0.62
3  0  0.31
4  0  0.16
```

Pruebas de póker

- Esta prueba mide la frecuencia de ciertas combinaciones de 5 números a la vez, (pero se puede cambiar por supuesto) basado en el juego de póker, como *aaaaa*, *aaaab*, *aaabb*, etc. Compara los resultados obtenidos en la muestra contra los valores teóricos, utilizando una prueba ji-cuadrada.
- Por ejemplo, en una “mano” de tamaño 3, hay tres posibilidades:
 - 1 Todos los dígitos son diferentes
 - 2 Todos son iguales
 - 3 Hay dos dígitos iguales

Las probabilidades de los eventos son:

- 1 $P(\text{Caso 1}) = (0.9)(0.8) = 0.72$
- 2 $P(\text{Caso 2}) = 0.01$
- 3 $P(\text{Caso 3}) = \binom{3}{2}(0.1)(0.9) = 0.27$

Ejemplo de aplicación

```
separa <- function(x){ #separa un número en sus dígitos componentes.
  w <- substring(as.character(x),3)
  if(nchar(w) < 4) for(i in 1:(4-nchar(w))) w <- paste(w, "0", sep = "")
  return(as.numeric(unlist(strsplit(w,""))))
}

tabla <- function(x,k){ # Hace una tabla de frecuencias de cada dígito
  tabla <- table(x)
  aa <- tabla[match(0:(k-1),as.numeric(names(tabla)))]
  aa <- ifelse(is.na(aa),0,aa)
  r <- aa[1:k]
  names(r) <- 0:(k-1)
  return(r)
}
```

Ejemplo de aplicación

```
prueba.poker<-function(vec){
  z <- round(vec,4) #redondeo a 4 decimales
  N <- length(z)
  ow <- options("warn") #apaga los warnings por un momento
  options(warn = -1)
  dim(z) <- c(N,1) #convierte el vector de datos a una matriz columna
  z1 <- apply(z, 1, separa) #separa cada numero aleatorio en los componentes
  z2 <- t(apply(z1, 2, tabla, k=10)) #crea una tabla de frecuencias
  # frecuencias de ceros unos, dos y tres de la tabla de frecuencias anterior,
  # para caracterizar los posibles juegos en cada "mano":
  # pachuca: hay 6 ceros y 4 unos siempre.
  # un par : hay 7 ceros y 1 dos y 2 unos.
  # dos pares: hay 2 dos, 8 ceros
  # una tercia: hay 1 tres y 1 uno, 8 ceros
  # un pokar: hay 1 cuatro y 9 ceros
  z1 <- apply(z2, 1, table)
  pachuca <- sum(unlist(lapply(z1,function(x){ifelse(x[1]==6,1,0)})))
  unpar <- sum(unlist(lapply(z1,function(x){ifelse(x[1]==7,1,0)})))
  dospar <- sum(unlist(lapply(z1,function(x){ifelse((x[1]==8)&(x[2]==2),1,0)})))
  tercia <- sum(unlist(lapply(z1,function(x){ifelse((x[1]==8)&(length(x)==3),1,0)})))
  pokar <- sum(unlist(lapply(z1,function(x){ifelse(x[1]==9,1,0)})))
  esp <- N*c(0.504, 0.432, 0.027, 0.036, 0.001)
  obs <- c(pachuca, unpar, dospar, tercia, pokar)
  prueba <- sum((obs-esp)^2/esp)
  pval <- 1 - pchisq(prueba,4)
  #names(obs) <- 0:9
  options(ow)
  return(list(cbind(Esperado=esp,Observado=obs),Estadistica=prueba,pval=round(pval,5)))
}
```

Ejemplo de aplicación

```
prueba.poker(runif(1000))
```

```
[[1]]
```

	Esperado	Observado
[1,]	504	523
[2,]	432	412
[3,]	27	27
[4,]	36	36
[5,]	1	2

```
$Estadistica
```

```
[1] 2.6
```

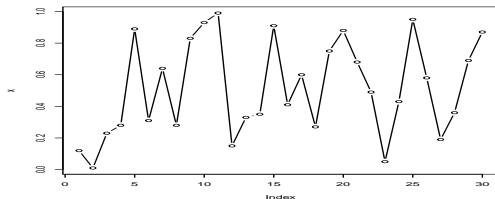
```
$pval
```

```
[1] 0.62
```


Pruebas para autocorrelación I

- Las pruebas de autocorrelación miden la dependencia entre subseries de una serie. Los números pueden estar relacionados de múltiples maneras:
 - pueden mostrar tendencias crecientes o decrecientes, o los números pueden estar intercalados.
 - pueden estar relacionados subseries, digamos cada i -ésimo número. En la serie que sigue, se puede ver que cada quinto número se tiene una observación grande:

```
x <- c( 0.12, 0.01, 0.23, 0.28, 0.89, 0.31, 0.64, 0.28, 0.83, 0.93,  
       0.99, 0.15, 0.33, 0.35, 0.91, 0.41, 0.60, 0.27, 0.75, 0.88,  
       0.68, 0.49, 0.05, 0.43, 0.95, 0.58, 0.19, 0.36, 0.69, 0.87)  
plot(x,type="b")
```



- La *autocorrelación* es una medida que nos dice cuánto puede depender una observación de otras observaciones generadas con el mismo proceso estocástico.

Pruebas para autocorrelación II

- Para una muestra X_1, \dots, X_n , la autocorrelación de rezago (lag) j se define como

$$\rho_j = \frac{\text{Cov}(X_i, X_{i+j})}{sd(X_i)sd(X_{i+j})} = \frac{\text{Cov}(X_i, X_{i+j})}{\text{Var}(X_1)},$$

si las variables tienen la misma distribución (o en general, el proceso es estacionario).

- Se pueden obtener estimadores a partir de una muestra de varias formas para la autocorrelación:

- $\text{Var}(X_1)$ se puede estimar con $s_n^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$.
- $\text{Cov}(X_i, X_{i+j})$ se puede estimar con $\hat{c}_j = \frac{\sum_{i=1}^{n-j} (X_i - \bar{X})(X_{i+j} - \bar{X})}{n-j}$, o bien con $\hat{c}'_j = \hat{c}_j \frac{n-j}{n}$

- En cualquiera de los casos anteriores,
 - i. $\hat{\rho}_j = \frac{\hat{c}_j}{s_n^2}$ es un estimador sesgado de ρ_j .
 - ii. Estimadores para diferentes rezagos estarán correlacionados, esto es: $\text{Cov}(\hat{\rho}_j, \hat{\rho}_k) \neq 0$.
 - iii. Si n es pequeña y j grande, entonces $\hat{\rho}_j$ es un estimador pobre para ρ_j .

Una prueba se puede basar en las autocorrelaciones para diferentes valores del rezago j . La hipótesis a probar es que $H_0 : \rho_j = 0$ para $j = 1, \dots, k$. Otra forma de ver esta prueba es que las observaciones forman una serie de *ruido blanco*.

Prueba adaptada para uniformes

Sean $u_1, \dots, u_n \sim \mathcal{U}(0, 1)$. Para una j dada, queremos probar:

$$H_0 : \rho_j = 0 \quad H_1 : \rho_j \neq 0$$

Como $E(u_i) = \frac{1}{2}$, $\text{Var}(u_i) = \frac{1}{12}$, entonces

$$\rho_j = \frac{c_j}{\sigma^2} = \frac{E(u_i u_{i+j}) - \frac{1}{4}}{\frac{1}{12}} = 12E(u_1 u_{1+j}) - 3$$

para cualquier i (la serie es estacionaria).

Estimamos $E(u_1 u_{1+j})$ con: $\frac{1}{h+1} \sum_{k=0}^h u_{1+kj} u_{1+(k+1)j}$ donde $h = \lfloor \frac{n-1}{j} \rfloor - 1$ ³.

h es el número de pares que se pueden formar, cuando las observaciones están espaciadas cada j observaciones.

³ $\lfloor x \rfloor = n$ si $n \leq x \leq n+1$, el entero más cercano a x que es menor o igual a x . En \mathbb{R} corresponde a la función floor.

Prueba adaptada para uniformes

De este modo, $\hat{\rho}_j = \frac{12}{h+1} \sum_{k=0}^h u_{1+kj} u_{1+(k+1)j} - 3$. Se puede probar con un poco de álgebra que $\text{Var}(\hat{\rho}_j) = \frac{13h+7}{(h+1)^2}$. Usando el teorema del límite central para sumas de variables aleatorias,

$$A_j = \frac{\hat{\rho}_j}{\sqrt{\text{Var}(\hat{\rho}_j)}} \sim \mathcal{N}(0, 1).$$

Aquí aplicamos una prueba estándar de variables aleatorias normales.

En `pruebas.r` se encuentra la función `prueba.correl` y tiene 2 argumentos: `init` que da el valor a partir del punto en donde se comenzaron a contar los rezagos, y `sig` que da el nivel de significancia escogido. El output de la función corresponde a los valores de la estadística de prueba A_j para diferentes j y los p -values obtenidos de la prueba. Adicionalmente, se genera una gráfica para visualizar los p -values.

```
> prueba.correl(x, init=1, sig=0.01)
```

Autocorrelaciones para ruido blanco

- Alternativamente, se puede calcular la función de autocorrelación para una serie de observaciones. en R, la función `acf` calcula la función de autocorrelación para varios rezagos y aplica la prueba descrita anteriormente.
- Debe quedar claro que siempre $\rho_0 = 1$ y que los límites de confianza que se muestran en la gráfica que se obtiene pueden dejar afuera 5 % de las observaciones.
- Si la serie muestra dependencia, entonces algunos valores de las autocorrelaciones pueden salir muy significativas.
- Adicionalmente, se puede calcular la función de autocorrelación parcial, `pacf` que complementa a la función `acf`.

Función de autocorrelación

Si una serie es débilmente estacionaria, entonces

$$\rho_k = \frac{Cov(y_t, y_{t-k})}{\sqrt{var(y_t)var(y_{t-k})}} = \frac{\gamma_k}{\gamma_0}.$$

La autocorrelación tiene las siguientes propiedades:

- $\rho_0 = 1$
- $-1 \leq \rho_k \leq 1$.
- $\{y_t\}$ no es serialmente correlacionada si $\rho_k = 0 \quad \forall k > 0$.

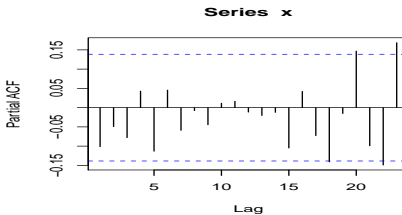
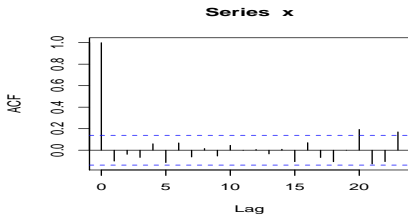
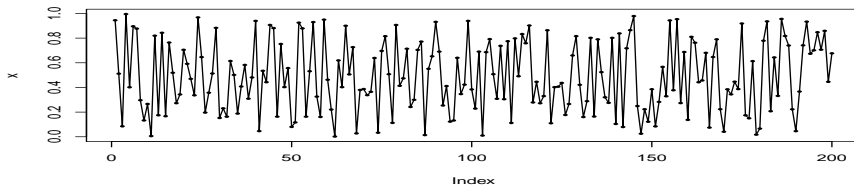
Para una serie estacionaria, el coeficiente de autocorrelación muestral se define como:

$$r_k = \frac{\sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

la gráfica de r_k vs. k es el correlograma de una serie temporal

Ejemplo

```
layout(matrix(c(1,1,2,3),nrow=2,byrow=T)); x <- runif(200)
plot(x, type = "o", pch = 16, cex = 0.6)
acf(x)
pacf(x)
```



Pruebas de autocorrelación: Bartlett

Se puede probar que $\hat{\rho}_1$ es un estimador consistente de ρ_1 . En particular, si $\{y_t\}$ es una serie de ruido blanco (iid con $\mu = 0$, $\sigma^2 < \infty$) entonces

$$\hat{\rho}_1 \sim N(0, 1/N)$$

En la práctica, se prueba la hipótesis $H_0 : \rho_1 = 0$ vs $H_a : \rho_1 \neq 0$ utilizando la estadística de prueba $t = \sqrt{N}\hat{\rho}_1$ que tiene distribución normal estándar, bajo la hipótesis nula. Lo mismo sucede para $\hat{\rho}_k$. Esta prueba se conoce como la **prueba de Bartlett** (no confundir con la prueba de Bartlett de homoscedasticidad).

Observación: cuando N es pequeña, digamos menor que 30, la prueba puede ser sesgada.

- Cuando la serie de tiempo es un ruido blanco, las propiedades de la función de autocorrelación son bien conocidas, y esto nos puede ayudar.
 - Los coeficientes de autocorrelación de una serie de ruido blanco se aproximan a una distribución normal con media 0 y varianza $\frac{1}{n}$, donde n es el número de observaciones en la serie. Así que 95 % de los coeficientes de autocorrelación deben estar entre $\pm 1.96/\sqrt{n}$, que son los límites críticos incluidos en las gráficas.
 - También las autocorrelaciones parciales deben ser cercanas a 0 cuando el modelo es un modelo de ruido blanco.

Pruebas de autocorrelación: Box-Pierce y Ljung-Box

La prueba de Box-Pierce (1970) prueba simultáneamente que varias autocorrelaciones son 0:

$$H_0 : \rho_1 = \dots = \rho_m = 0 \quad vs. \quad \rho_i \neq 0 \text{ para alguna } i$$

La estadística de prueba en este caso es $Q^*(m) = N \sum_{i=1}^m \hat{\rho}_i^2$. Se puede demostrar que la distribución asintótica de $Q^*(m)$ es una ji-cuadrada con m grados de libertad (χ_m^2)

La prueba de Ljung-Box (1978) modifica $Q^*(m)$ para incrementar el poder estadístico de la prueba cuando se tienen muestras pequeñas. En este caso, se considera

$$Q(m) = N(N+2) \sum_{i=1}^m \frac{\hat{\rho}_i^2}{N-i}$$

En la práctica una buena elección de m es tomar $m = \log(N)$.

Ejemplos I

Llevar a cabo las pruebas de Bartlett, Box-Pierce y Ljung-Box para los siguientes datos:

```
set.seed(1)
u <- runif(100)
u
```

```
[1] 0.266 0.372 0.573 0.908 0.202 0.898 0.945 0.661 0.629 0.062 0.206 0.177 0.687 0.384 0.770 0.498 0.718 0.992 0.380
[20] 0.777 0.935 0.212 0.652 0.126 0.267 0.386 0.013 0.382 0.870 0.340 0.482 0.600 0.494 0.186 0.827 0.668 0.794 0.108
[39] 0.724 0.411 0.821 0.647 0.783 0.553 0.530 0.789 0.023 0.477 0.732 0.693 0.478 0.861 0.438 0.245 0.071 0.099 0.316
[58] 0.519 0.662 0.407 0.913 0.294 0.459 0.332 0.651 0.258 0.479 0.766 0.084 0.875 0.339 0.839 0.347 0.334 0.476 0.892
[77] 0.864 0.390 0.777 0.961 0.435 0.713 0.400 0.325 0.757 0.203 0.711 0.122 0.245 0.143 0.240 0.059 0.642 0.876 0.779
[96] 0.797 0.455 0.410 0.811 0.605
```

Solución.

Ejemplos II

```
#Prueba de Bartlett:  
rho1 <- acf(u,1,plot=F)$acf[2]  
bt <- sqrt(length(u))*rho1  
1-pnorm(bt)/2 #p-value  
  
[1] 0.72  
  
#Prueba de Box-Pierce:  
Box.test(u,lag=3,type="Box-Pierce")  
  
Box-Pierce test  
  
data: u  
X-squared = 3, df = 3, p-value = 0.4  
  
#Prueba de Ljung-Box:  
Box.test(u,lag=3,type="Ljung-Box")  
  
Box-Ljung test  
  
data: u  
X-squared = 3, df = 3, p-value = 0.4
```



Función de autocorrelación parcial

- Las *autocorrelaciones parciales* se usan para medir el grado de asociación entre y_t y y_{t-k} , cuando se ha eliminado el efecto de los rezagos intermedios $1, 2, 3, \dots, k-1$.
- El coeficiente de correlación parcial de orden k se denota por α_k y se calcula haciendo la regresión de y_t contra los rezagos y_{t-1}, \dots, y_{t-k} :

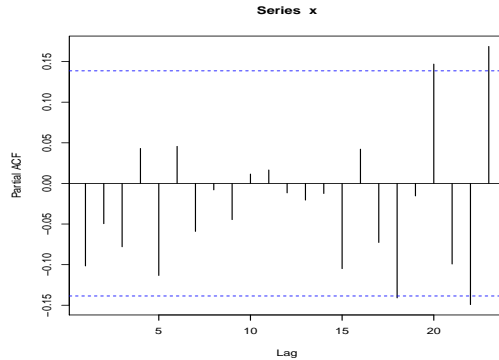
$$y_t = b_0 + b_1 y_{t-1} + \dots + b_k y_{t-k}$$

A esta regresión se le llama *autorregresión*, porque es y contra si misma.

- El valor α_k es el coeficiente b_k de esta regresión.
- La función de autocorrelación parcial se grafica: α_k vs. k .

Ejemplo

```
pacf(x)
```



Interpretación de las gráficas

- Atributos de las gráficas:
 - En ambas gráficas, acf y pacf, se pueden observar dos líneas azules, basadas en los límites discutidos: estas sirven para indicar qué correlaciones son significativas (las que rebasen las líneas son importantes).
 - La cola de la acf se acerca a 0 con un decaimiento sinusoidal.
 - La pacf sólo tiene dos rezagos significativos.
- En el estudio de series de tiempo, estas dos gráficas nos puede ayudar a identificar un modelo de tipo autorregresivo y de promedios móviles que sirve para estimar la serie de tiempo (modelos ARIMA de Box y Jenkins).

Otras baterías de pruebas para números aleatorios

- Como diferentes pruebas son sensibles a diferentes tipos de desviaciones de la hipótesis nula de uniformidad e independencia, se requiere un conjunto de pruebas que recorra el espacio de hipótesis alternativas.
- Algunos ejemplos de baterías de pruebas que son populares son los siguientes
 - Fishman & Moore (1982, 1986): pruebas de bondad de ajuste sobre transformaciones de la muestra.
 - Vattulainen, Ala-Nissila & Kankaala (1994, 1995): basados en modelos físicos.
 - **DIEHARD tests** de Marsaglia (1985, 1995): incluye 18 pruebas de bondad de ajuste.
 - **DIEHARDer tests** de Robert G. Brown (2003): prueba generadores, no conjuntos de datos. Combina Marsaglia y NIST.
 - **NIST Test Suite** (2000)
 - **TestU01** (L'Ecuyer, 1985) (incluye DIEHARD y NIST): son aproximadamente 60 pruebas.