

PROBANDO GENERADORES DE NUMEROS ALEATORIOS

Es importante asegurarse de que el generador usado produzca una secuencia suficientemente aleatoria. Para esto se somete el generador a pruebas estadísticas. Si no pasa una prueba, podemos asumir que el generador es malo. Pasar una prueba es una condición necesaria pero no suficiente. Un generador puede pasar una prueba y luego no pasarla si se usa otra semilla u otro segmento del ciclo.

Describiremos pruebas para números aleatorios uniformemente distribuidos, aunque muchas de las pruebas también pueden ser usadas para probar variables aleatorias.

Lo primero en la prueba de un generador, es graficar y observar las distribuciones de un histograma y la frecuencia acumulada. Para el generador programado anteriormente (el cual usaremos para ilustrar todas las pruebas):

$$x_n = 7^5 x_{n-1} \pmod{(2^{31} - 1)}$$

y usando $x_0 = 1$, se obtiene la siguiente tabla de frecuencias, el histograma, y gráfico de la frecuencias acumuladas, a partir de una secuencia de 200 de números aleatorios:

Tabla de Frecuencias

<.05	<.10	<.15	<.20	<.25	<.30	<.35	<.40	<.45	<.50	<.55	<.60	<.65	<.70	<.75	<.80	<.85	<.90	<.95	<1.0
10	11	11	8	9	11	6	10	10	15	12	8	8	12	8	9	12	8	13	9

Histograma de Frecuencias Relativas

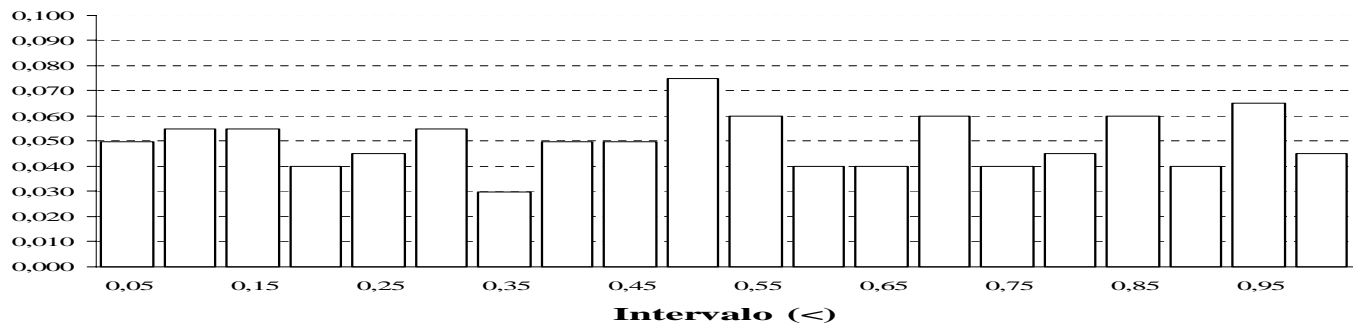
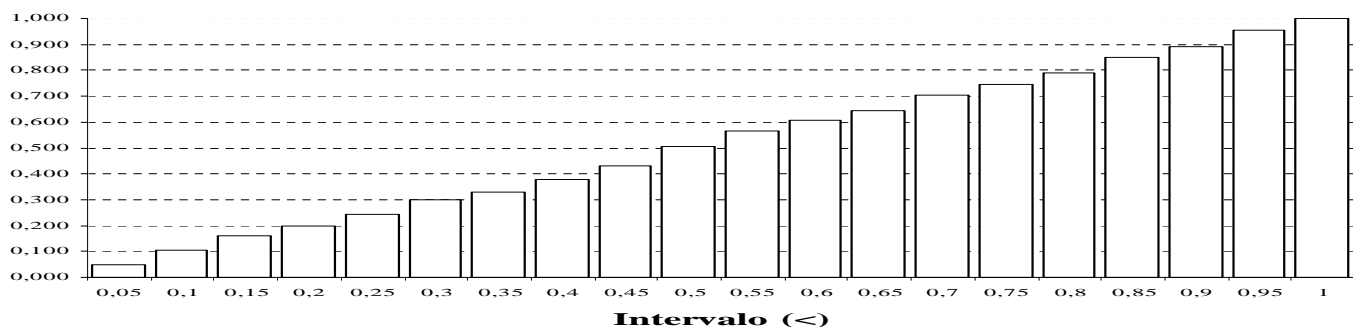


Gráfico de Frecuencias Acumuladas



Vemos que los gráficos son los esperados. Todas las frecuencias en el histograma son aproximadamente 0.05 y el gráfico de las frecuencias acumuladas es aproximadamente una línea recta de 45°.

I. PRUEBA DE CHI-CUADRADO

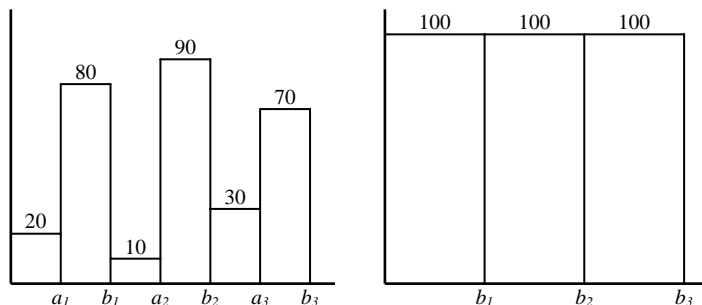
Esta es la prueba más comúnmente usada. En general, puede ser usada para cualquier distribución. A partir de un histograma, se comparan las frecuencias observadas con las frecuencias obtenidas de la distribución específica (frecuencias esperadas). Si el histograma tiene k celdas o intervalos, y O_i y E_i son las frecuencias observadas y esperadas respectivamente para la i -ésima celda, la prueba consiste en calcular

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Si el ajuste es exacto, χ_0^2 es cero, pero por aleatoriedad no lo será. Se puede demostrar que χ_0^2 tiene distribución chi-cuadrado con $k-1$ grados de libertad. Se aceptara que los datos tienen la distribución en prueba con un nivel de significancia de α^* si $\chi_0^2 < \chi_{[\alpha; k-1]}^2$.

Para los datos anteriores, todas las frecuencias esperadas son de $200/20=10$. El valor de χ_0^2 es 8,80 y la $\chi_{[0.10; 19]}^2 = 27,2$; por lo tanto no hay evidencia de que los números no son uniformes a un nivel de $\alpha=0.10$.

Estrictamente hablando, la prueba de chi-cuadrado esta diseñada para distribuciones discretas y para muestras grandes. Para distribuciones continuas la chi-cuadrado es solo una aproximación y el nivel de significación se aplica solo si $n \rightarrow \infty$. Para muestras finitas el nivel de significación es algo menor. Si en un caso particular una celda contiene menos de 5 observaciones esperadas, diversos autores recomiendan que se combinen celdas adyacentes a fin de que cada celda tenga al menos 5 observaciones esperadas. Uno de los problemas de esta prueba es la selección de los límites de las celdas. Esto puede afectar las conclusiones y no hay reglas exactas para seleccionar su tamaño. Para darnos cuenta de la situación, consideremos los dos histogramas siguientes, ambos construidos a partir de los mismos datos. En el de la izquierda difícilmente se puede llegar a la conclusión de que los datos provienen de una población uniforme. Si vamos al de la derecha, en donde tenemos los mismos datos pero reagrupados, la historia es otra y en este caso no tenemos dudas de la uniformidad; la prueba de chi-cuadrado da un ajuste perfecto.



* Recordemos que α es la probabilidad de cometer Error de Tipo I: rechazar H_0 cuando es verdadera.

Para ayudar en la determinación del número de intervalos se puede tener en cuenta lo siguiente:

- Si la distribución en estudio es discreta, cada valor de la variable aleatoria debería ser una clase, a no ser que sea necesario combinar celdas adyacentes para cumplir con la frecuencia esperada mínima por celda. En este caso $p_i = p(x_i) = P(X = x_i)$, o p_i se determina sumando las probabilidades de celdas adyacentes.
- Si la distribución es continua la siguiente tabla puede ser usada de guía:

Tamaño de la muestra n	Número de intervalos k
20	No use la prueba de chi-cuadrado
50	5 a 10
100	10 a 20
>100	\sqrt{n} a $n/5$

Notese que para el ejemplo anterior con 200 observaciones, el número de intervalos debería estar entre $\sqrt{200}$ y $200/5$ o entre 15 y 40. Hay 20 intervalos, así que se está dentro del rango sugerido.

Otro problema es el hecho de que la frecuencia esperada está dividiendo, lo que implica que errores en celdas con frecuencias esperadas pequeñas afectan más el valor de χ_0^2 que errores en celdas con frecuencias esperadas grandes.

II. PRUEBA DE KOLMOGOROV-SMIRNOV

La prueba K-S nos permite decidir si una muestra de n observaciones es de una distribución continua particular. Se basa en que la diferencia entre la FDA (Función de Distribución Acumulada) observada $S_n(x)$ y la FDA esperada $F_X(x)$ debe ser pequeña. $S_n(x)$ viene dada por:

$$S_n(x) = \frac{\text{número de observaciones} \leq x}{n}$$

Los símbolos D^+ y D^- son usados para denotar las desviaciones observadas máximas sobre y bajo la FDA esperada en una muestra de tamaño n :

$$D^+ = \max_x [S_n(x) - F_X(x)] \quad D^- = \max_x [F_X(x) - S_n(x)]$$

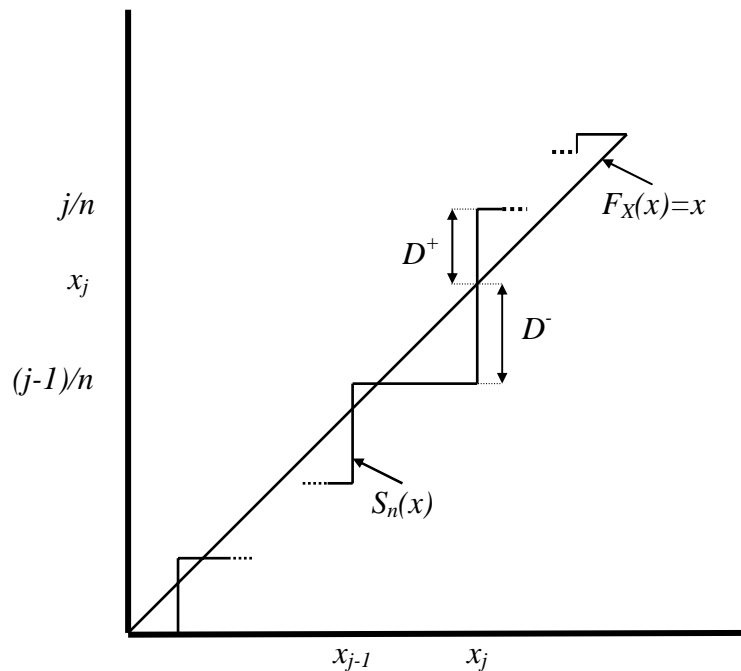
D^+ mide la desviación máxima cuando la FDA observada está sobre la FDA esperada, y D^- mide la desviación máxima cuando la FDA observada está bajo la FDA esperada. Si los valores de D^+ y D^- son menores que $D_{[1-\alpha; n]}$ entonces se dice que las observaciones provienen de la distribución respectiva con un nivel de significación de α .

Un error común calculando D^- consiste en encontrar el máximo de $F_X(x_j) - S_n(x_j)$. Esto es incorrecto ya que S_n consiste de un segmento horizontal a $S_n(x_j)$ en el intervalo $[x_j, x_{j+1})$ y el máximo ocurre justo antes de

x_{j+1} y la diferencia correcta es $F_X(x_{j+1}) - S_n(x_j)$. Vea el grafico siguiente en donde se muestra como se hacen los cálculos y se asume que ambos máximos D^+ y D^- caen en x_j , cosa que no tiene que ser así.

Para números aleatorios uniformes entre 0 y 1, la FDA esperada es $F_X(x) = x$. Para probar uniformidad ordenamos la muestra $\{x_1, x_2, \dots, x_n\}$ tal que $x_{j-1} < x_j$ y calculamos

$$D^+ = \max_j \left(\frac{j}{n} - x_j \right) \quad D^- = \max_j \left(x_j - \frac{j-1}{n} \right)$$



Comparando con la D podemos tomar la decisión.

Ejemplo

Para el generador que hemos venido usando, una muestra de tamaño 30 partiendo de $x_0 = 1$ es:

0.000008, 0.131538, 0.755605, 0.458650, 0.532767, 0.218959, 0.047045, 0.678865,
 0.679296, 0.934693, 0.383502, 0.519416, 0.830965, 0.034572, 0.053462, 0.529700,
 0.671149, 0.007698, 0.383416, 0.066842, 0.417486, 0.686773, 0.588977, 0.930436,
 0.846167, 0.526929, 0.091965, 0.653919, 0.415999, 0.701191.

de donde obtenemos los siguientes valores: $D^+ = 0.14137$ y $D^- = 0.08342$. $D_{[0.1; 30]} = 0.22$ y por lo tanto decimos que no hay evidencia de que los números no sean uniformes.

La prueba K-S esta diseñada para distribuciones continuas, muestras pequeñas, y usas todas las observaciones sin hacer agrupaciones haciendo mejor uso de los datos que la chi-cuadrado.

III. PRUEBA DE CORRELACIÓN SERIAL

Un método para probar si hay dependencia entre dos variables es calculando su covarianza. Si esta es distinta de cero, entonces son dependientes. En inverso no es cierto; si la covarianza es cero no implica que sean independientes.

Dada una secuencia de números aleatorios uniformes entre 0 y 1 ($U(0,1)$), se puede calcular la covarianza entre números k -distantes, es decir, entre x_i y x_{i+k} . Esto es llamado **autocovarianza con desplazamiento k** denotada como R_k y dada por:

$$R_k = \frac{1}{n-k} \sum_{i=1}^{n-k} \left(U_i - \frac{1}{2} \right) \left(U_{i+k} - \frac{1}{2} \right)$$

donde U_i es el i -ésimo número aleatorio uniforme de la secuencia.

Para n grande, R_k se distribuye normalmente con media 0 y varianza $1/[144(n-k)]$. El intervalo de confianza para la autocovarianza al $100(1-\alpha)\%$ es

$$R_k \mp z_{1-\alpha/2} / (12\sqrt{n-k})$$

Si este intervalo no incluye el cero, podemos decir que la secuencia tiene correlación significativa. Esto se aplica solo para $k \geq 1$. Si $k = 0$, R_0 es la varianza de la muestra que se espera sea $1/12$ para una secuencia independiente idénticamente distribuida (IID) de $U(0,1)$.

Ejemplo

Para la muestra anterior con 10000 números:

Distanciamiento k	Autocovarianza R_k	Intervalo de Confianza al 95%	
		Limite inferior	Limite superior
1	-0.0000383	-0.0016701	0.0015934
2	-0.0010171	-0.0026489	0.0006147
3	-0.0004891	-0.0021211	0.0011428
4	-0.0000325	-0.0016645	0.0015995
5	-0.0005311	-0.0021632	0.0011009
6	-0.0012766	-0.0029087	0.0003556
7	-0.0003854	-0.0020176	0.0012469
8	-0.0002072	-0.0018395	0.0014251
9	0.0010313	-0.0006011	0.0026637

Todos los intervalos incluyen el cero y podemos asumir todas las covarianzas como estadísticamente no significativas al nivel de confianza $\alpha=0.05$ o $Z_{0.975}=1.958$

IV. PRUEBAS DE DOS NIVELES

En las pruebas anteriores, si la muestra es muy pequeña, los resultados son locales y no se aplican a todo el ciclo (no son globales). Similarmente, resultados globales pueden que no se apliquen localmente ya que puede haber muy poca aleatoriedad en ciertos segmentos del ciclo.

Para resolver este problema se propone usar pruebas de dos niveles. Por ejemplo, usar chi-cuadrado en n muestras de tamaño k y después usar chi-cuadrado en las n chi-cuadrados obtenidas. Similarmente se puede hacer con la prueba K-S.

Usando pruebas de este tipo se han encontrado segmentos “no aleatorios” en secuencias que de otra forma son aleatorias.

V. UNIFORMIDAD K-DIMENSIONAL O K-DISTRIBUTIVA

Las pruebas anteriores aseguran uniformidad en una dimensión. El concepto de uniformidad puede ser extendido a k dimensiones.

Supongamos que u_n es el n -ésimo número en una secuencia distribuida uniformemente entre 0 y 1. Dados dos números reales a_1 y b_1 , también entre 0 y 1 y tal que $b_1 > a_1$, la probabilidad de que u_n este en el intervalo $[a_1, b_1)$ es $b_1 - a_1$:

$$P(a_1 \leq u_n < b_1) = b_1 - a_1 \quad \forall b_1 > a_1$$

Esta es la propiedad 1-distributiva de u_n .

La 2-distributividad es una generalización de esta propiedad en dos dimensiones y requiere que un par de valores u_{n-1} y u_n satisfagan la siguiente condición:

$$P(a_1 \leq u_{n-1} < b_1 \text{ y } a_2 \leq u_n < b_2) = (b_1 - a_1)(b_2 - a_2) \quad \forall b_1 > a_1 \text{ y } \forall b_2 > a_2$$

Una secuencia es k -distributiva si:

$$P(a_1 \leq u_n < b_1, \dots, a_k \leq u_{n+k-1} < b_k) = (b_1 - a_1) \dots (b_k - a_k) \quad \forall b_i > a_i \quad i = 1, 2, \dots, k$$

Note que una secuencia k -distributiva es $(k-1)$ -distributiva, pero el inverso no es cierto. Una secuencia puede ser uniforme en una dimensión inferior y no en una superior. Dado un grupo de generadores, es preferible aquel que produzca más uniformidad en la mayor dimensión. A continuación veremos una forma de chequear k -distributividad: prueba serial.

Antes de conducir estas pruebas, es conveniente chequear si la secuencia es uniforme en dos dimensiones graficando pares solapados sucesivos: (x_{n-1}, x_n) , (x_n, x_{n+1}) .

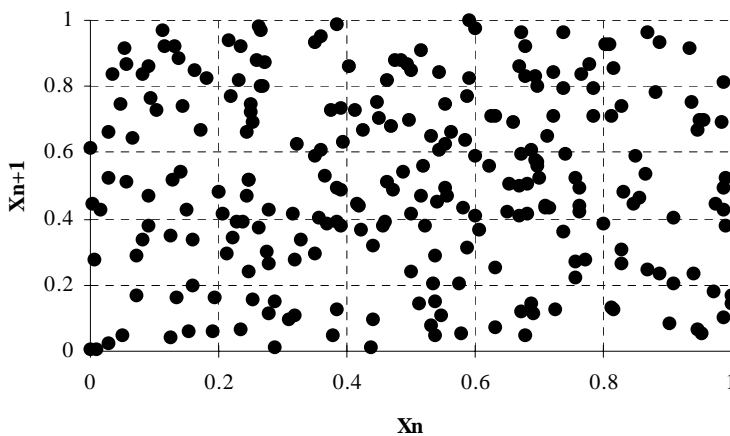
VI. PRUEBA SERIAL

Esta prueba es usada para probar uniformidad en dos o más dimensiones. En dos dimensiones, se divide el espacio entre 0 y 1 en K^2 celdas de igual área. Si tenemos una muestra de tamaño n , podemos construir $n/2$ pares no solapados (x_1, x_2) , (x_3, x_4) , ..., y contar los puntos que caen en cada celda. Idealmente se esperan

$\frac{n/2}{K^2}$ puntos en cada celda. Se puede usar la chi-cuadrado para encontrar la desviación entre lo observado y lo esperado. Los grados de libertad son K^2-1 . Es fácil ver como se puede extender la prueba a k -dimensiones.

Ejemplo:

Para el generador ejemplo, usando una muestra de tamaño 500, y dividiendo en $5^2 = 25$ celdas que dan 250 pares no solapados, obtenemos:



y las siguientes frecuencias:

11	12	10	11	7
8	10	13	9	8
9	11	13	21	10
7	14	10	6	8
10	9	8	6	9

Con 250 pares esperamos 10 observaciones por celda. El valor de χ_0^2 es 23,2 y la $\chi_{[0.10; 24]}^2 = 33,2$; por lo tanto aceptamos que los números son uniformes en dos dimensiones a un nivel de $\alpha=0.10$.

No se deben usar pares solapados. Si se usan pares solapados, el número de puntos por celda no es independiente y no se puede usar la prueba chi-cuadrado. La dependencia entre números sucesivos aparece como no-uniformidad en dimensiones más grandes. Por ejemplo, si números sucesivos tienen correlación de primer orden negativa, es muy probable un valor alto x_n sea seguido de un valor bajo x_{n+1} . Si graficamos los pares no solapados, los puntos tienden a concentrarse derecha-y-abajo e izquierda-y-arriba y no se pasará la prueba.

VII.PRUEBAS DE RACHAS

Consideremos la siguiente secuencia de números:

0.00001 0.00770 0.03457 0.04704 0.04746 0.05346 0.06684 0.09196 0.13154 0.21896
0.26245 0.32823 0.36534 0.38342 0.38350 0.41600 0.41749 0.45865 0.51942 0.52693
0.52970 0.53277 0.58898 0.63264 0.65392 0.67115 0.67886 0.67930 0.68677 0.70119
0.73608 0.75561 0.75641 0.76220 0.83097 0.84617 0.91032 0.93044 0.93469 0.99104

difícilmente se puede decir que esta secuencia es aleatoria (los número están ordenados), sin embargo, pasan las pruebas de uniformidad chi-cuadrado y Kolmogorov-Smirnov:

Para la prueba chi-cuadrado tenemos las siguientes frecuencias observadas y esperadas:

Intervalo	O_i	E_i
$0.000 \leq x < 0.200$	9	8
$0.200 \leq x < 0.400$	6	8
$0.400 \leq x < 0.600$	8	8
$0.600 \leq x < 0.800$	11	8
$0.800 \leq x < 1.000$	6	8
Total:	40	40

Con $\chi_0^2 = 2,25$ y $\chi_{[0.10;4]}^2 = 7,78$, y para la prueba de Kolmogorov-Smirnov $D^+ = 0.10816$, $D^- = 0.06942$ y $D_{[0.1;40]} = 1.22/\sqrt{40} = 0.19$ y vemos que en ambos casos no hay evidencia para rechazar la uniformidad de la secuencia. Esto es fácil de entender al ver que ninguna de las dos pruebas le presta atención al orden de los números. Las pruebas de rachas, en donde una **racha** es una secuencia de eventos de cierto tipo, sí toman en consideración la forma como se dan los números en la secuencia.

Rachas Crecientes y Decrecientes

Una racha creciente es aquella en que un número esta seguido por un número mayor, mientras que en la decreciente un número esta seguido por un número menor. Consideremos la siguiente secuencia, que proviene del generador que hemos usado de ejemplo y que si los ordenamos da la secuencia al comienzo de la sección:

0.00001 0.13154 0.75561 0.45865 0.53277 0.21896 0.04704 0.67886 0.67930 0.93469
0.38350 0.51942 0.83097 0.03457 0.05346 0.52970 0.67115 0.00770 0.38342 0.06684
0.41749 0.68677 0.58898 0.93044 0.84617 0.52693 0.09196 0.65392 0.41600 0.70119
0.91032 0.76220 0.26245 0.04746 0.73608 0.32823 0.63264 0.75641 0.99104 0.36534

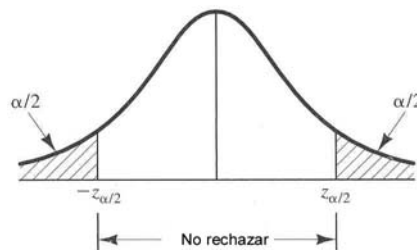
Si usamos signos más y menos para identificar si un número esta en una racha creciente o decreciente respectivamente, tenemos:

+	+	-	+	-	-	+	+	+	-
+	+	-	+	+	+	-	+	-	+
+	-	+	-	-	-	+	-	+	+
-	-	-	+	-	+	+	+	-	

Sea N ($N > 20$) la longitud de la secuencia y a el número de rachas en la misma, si los números efectivamente son aleatorios, a se distribuye normalmente y:

$$\mu_a = \frac{2N - 1}{3}, \quad \sigma_a^2 = \frac{16N - 29}{90}, \quad Z_0 = \frac{a - \mu_a}{\sigma_a} \sim N(0,1)$$

y si $-Z_{\alpha/2} \leq Z_0 \leq Z_{\alpha/2}$ no hay evidencia para rechazar la hipótesis de independencia de los números, como se muestra en la siguiente figura.



Nótese que el número máximo de rachas es $N-1$ (con el último número de la secuencia no se puede decidir a que tipo de racha pertenece por ser precisamente el último), mientras que el mínimo es 1.

Para nuestro ejemplo $a = 24$, $\mu_a = 26.33$, $\sigma_a^2 = 6.79$, $Z_0 = -0.896$, $Z_{0.05} = 1.645$ y por lo tanto no rechazamos la hipótesis de independencia.

Rachas Bajo y Sobre la media

Una racha bajo la media es aquella en que el número es menor o igual a 0.5 (la media de la uniforme), mientras que es sobre si el número es mayor a 0.5. Consideremos la misma secuencia anterior y si usamos signos más y menos para identificar si un número esta sobre o bajo la media, tenemos:

-	-	+	-	+	-	-	+	+	+
-	+	+	-	-	+	+	-	-	-
-	+	+	+	+	+	-	+	-	+
+	+	-	-	+	-	+	+	+	-

Sean n_1 y n_2 la cantidad de números sobre y debajo de la media respectivamente, $N = n_1 + n_2$ la longitud de la secuencia y b el número de rachas en la misma, si los números efectivamente son aleatorios y n_1 o n_2 es mayor que 20, b se distribuye normalmente y:

$$\mu_b = \frac{2n_1n_2}{N} + \frac{1}{2}, \quad \sigma_b^2 = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N-1)}, \quad Z_0 = \frac{b - \mu_b}{\sigma_b} \sim N(0,1)$$

y si $-Z_{\alpha/2} \leq Z_0 \leq Z_{\alpha/2}$ no hay evidencia para rechazar la hipótesis de independencia de los números. Ahora el número máximo de rachas es N y el mínimo es 1.

Para nuestro ejemplo $b = 21$, $n_1 = 22$, $n_2 = 18$, $\mu_b = 20.30$, $\sigma_b^2 = 9.54$, $Z_0 = 0.227$, $Z_{0.10/2} = 1.645$ y por lo tanto no rechazamos la hipótesis de independencia.

Longitud de las rachas

No es importante solamente el tipo de rachas que se tiene, sino también cuanto es su longitud. Muchas rachas muy pequeñas o rachas muy largas producen serias sospechas. Sea Y_i la cantidad de rachas de longitud i en una secuencia de N números.

Para una secuencia independiente de rachas crecientes/decrecientes los valores esperados de Y_i están dados por:

$$E(Y_i) = \frac{2}{(i+3)!} [N(i^2 + 3i + 1) - (i^3 + 3i^2 - i - 4)] \quad \text{si } i \leq N - 2$$

$$E(Y_i) = \frac{2}{N!} \quad \text{si } i = N - 1$$

Para las rachas sobre y bajo la media tenemos:

$$E(Y_i) = \frac{Nw_i}{E(I)}$$

$$w_i = \left(\frac{n_1}{N}\right)^i \left(\frac{n_2}{N}\right) + \left(\frac{n_1}{N}\right) \left(\frac{n_2}{N}\right)^i$$

$$E(I) = \frac{n_1}{n_2} + \frac{n_2}{n_1}$$

$$E(A) = \frac{N}{E(I)}$$

Donde w_i es la probabilidad aproximada de que una racha tenga longitud i , $E(I)$ es la longitud promedio esperada de las rachas y $E(A)$ es el número esperado aproximado de rachas (de cualquier longitud) en la secuencia.

Con esto podemos aplicar la prueba de chi-cuadrado usando

$$\chi_0^2 = \sum_{i=1}^L \frac{(O_i - E(Y_i))^2}{E(Y_i)}$$

donde $L = N-1$ para rachas creciente/decrecientes y $L = N$ para rachas sobre/bajo la media. Comparamos y se acepta que los datos son independientes sí, con un nivel de significancia de α , $\chi_0^2 < \chi_{[\alpha; L-1]}^2$.

Recuerde que para esta prueba diversos autores recomiendan que las observaciones esperadas por intervalo deben ser mayores a 5 y si esto no se da se deben agrupar intervalos.

Con la secuencia que hemos venido estudiando, para las rachas crecientes/decrecientes tenemos:

Rachas de longitud i	Frecuencia Observada	Frecuencia Esperada
1	14	16.750
2	5	7.100
≥ 3	5	2.483

y como la última clase tiene un valor esperado menor que 5 la agrupamos con la segunda. Obsérvese que no hay rachas de longitud mayor que 3, por lo tanto la frecuencia esperada de la clase " ≥ 3 " se calcula como $(\mu_a - 16.750 - 7.100)$. Reagrupando da:

Rachas de longitud i	Frecuencia Observada	Frecuencia Esperada
1	14	16.750
≥ 2	10	9.583

con $\chi_0^2 = 0.470$ y $\chi_{[0.10;1]}^2 = 2.71$, llevándonos a no rechazar la hipótesis de independencia.

Para las rachas sobre/debajo de la media, la tabla de frecuencias que obtenemos es:

Rachas de longitud i	Frecuencia Observada	Frecuencia Esperada
1	10	9.704
2	6	4.852
3	3	2.450
4	1	1.249
≥ 5	1	1.348

y nuevamente hay que reagrupar intervalos dado que tenemos frecuencias esperadas menores a 5. También, dado que no hay rachas de longitud mayor que 5, la frecuencia esperada de la clase " ≥ 5 " se calcula como $(E(A) - \text{suma de las frecuencias esperadas anteriores})$. Reagrupando da:

Rachas de longitud i	Frecuencia Observada	Frecuencia Esperada
1	10	9.704
≥ 2	11	9.900

con $\chi_0^2 = 0.131$ y $\chi_{[0.10;1]}^2 = 2.71$, llevándonos a no rechazar la hipótesis de independencia.

Nuestros 40 números aleatorios del generador $x_n = 7^5 x_{n-1} \bmod (2^{31} - 1)$ con $x_0 = 1$ pasan todas las pruebas, inclusive la de autocorrelación serial:

Distanciamiento k	Autocovarianza R_k	Intervalo de Confianza al 90%	
		Limite inferior	Limite superior
1	0.0137667	-0.0123608	0.0398943
2	-0.0122781	-0.0387472	0.0141911
3	-0.0115585	-0.0383829	0.0152659
4	-0.0093449	-0.0365394	0.0178495
5	-0.0112315	-0.0388117	0.0163487
6	0.0039861	-0.0239967	0.0319690
7	0.0243485	-0.0040552	0.0527521
8	-0.0012593	-0.0301034	0.0275847
9	-0.0045192	-0.0338248	0.0247864

VIII.PRUEBA DE BRECHAS

Se usa para determinar si los intervalos (brechas) entre la ocurrencia del mismo dígito son o no aleatorios. El siguiente ejemplo ilustra las brechas asociadas al dígito 3.

4, 1, 3, 5, 1, 7, 2, 8, 2, 0, 7, 9, 1, 3, 5, 2, 7, 9, 4, 1, 6, 3
3, 9, 6, 3, 4, 8, 2, 3, 1, 9, 4, 4, 6, 8, 4, 1, 3, 8, 9, 5, 5, 7
3, 9, 5, 9, 8, 5, 3, 2, 2, 3, 7, 4, 7, 0, 3, 6, 3, 5, 9, 9, 5, 5
5, 0, 4, 6, 8, 0, 4, 7, 0, 3, 3, 0, 9, 5, 7, 9, 5, 1, 6, 6, 3, 8
8, 8, 9, 2, 9, 1, 8, 5, 4, 4, 5, 0, 2, 3, 9, 7, 1, 2, 0, 3, 6, 3

Hay dieciocho 3 y por lo tanto tenemos 17 brechas asociadas a este dígito. La primera brecha tiene una longitud de 10 y su probabilidad, asumiendo independencia, viene dada por:

$$P(\text{brecha de } 10) = P(\text{no}3)P(\text{no}3)P(\text{no}3)P(\text{no}3)P(\text{no}3)P(\text{no}3)P(\text{no}3)P(\text{no}3)P(\text{no}3)P(\text{no}3)P(3) = (0.9)^{10}(0.1)$$

y la distribución teórica vendría dada por:

$$P(\text{brecha} \leq x) = F(x) = 0.1 \sum_{n=0}^x (0.9)^n = 1 - 0.9^{x+1}$$

Para aplicar la prueba a números aleatorios asociamos los dígitos 0, 1, ... a los intervalos [0, 0.1), [0.1, 0.2), ...

Para el ejemplo anterior con 110 números, tenemos 100 brechas (los números menos la cantidad de dígitos distintos) y podemos comparar la distribución observada contra la teórica usando la prueba de Kolmogorov-Smirnov. Para $\alpha=0.05$, tenemos que el valor crítico es:

$$D_{0.05} = \frac{1.36}{\sqrt{100}} = 0.136$$

Longitud de brecha	Frecuencia	Frecuencia Relativa	Frecuencia Relativa Acumulada S(x)	F(x)	F(x)-S(x)
0-3	35	0,35	0,35	0,3439	0,0061
4-7	22	0,22	0,57	0,5695	0,0005
8-11	17	0,17	0,74	0,7176	0,0224
12-15	9	0,09	0,83	0,8147	0,0153
16-19	5	0,05	0,88	0,8784	0,0016
20-23	6	0,06	0,94	0,9202	0,0198
24-27	3	0,03	0,97	0,9477	0,0223
28-31	0	0,00	0,97	0,9657	0,0043
32-35	0	0,00	0,97	0,9775	0,0075
36-39	2	0,02	0,99	0,9852	0,0048
40-43	0	0,00	0,99	0,9903	0,0003
44-47	1	0,01	1,00	0,9936	0,0064

De la tabla observamos que la diferencia maxima entre la distribución observada (S(x)) y la teorica (F(x)) es 0.0224 que es inferior al valor critico (0.136) por lo que no podemos rechazar la hipótesis de independencia de los digitos.

IX. PRUEBA POKER

Esta es una prueba de independencia basada en la frecuencia con que ciertos digitos se repiten en una serie de números. Su nombre se debe al popular juego de cartas Poker. Consideremos la siguiente serie de números con una repetición inusual de digitos:

0.255, 0.577, 0.331, 0.414, 0.828, 0.909, 0.033, 0.010

En cada caso aparecen uno de los tres digitos repetido y las posibilidades para este caso son:

- Los tres digitos distintos
- Los tres digitos iguales
- Un par de digitos iguales

y las probabilidades asociadas son:

$$P(\text{todos distintos}) = P(\text{segundo distinto del primero})P(\text{tercero distinto del segundo}) = (0.9)(0.8) = 0.72$$

$$P(\text{todos iguales}) = P(\text{segundo igual al primero})P(\text{tercero igual al segundo}) = (0.1)(0.1) = 0.01$$

$$P(\text{un par}) = 1 - 0.72 - 0.01 = 0.27$$

Supongamos una secuencia de 1000 números aleatorios en donde se analizan los tres primeros digitos y se tiene que 680 tienen los 3 digitos distintos, 289 contienen exactamente un par y 31 tienen todos iguales. Los calculos respectivos usando la prueba Chi-Cuadrado son:

Caso	Frecuencias Observadas O_i	Frecuencias Esperadas E_i	$\frac{(O_i - E_i)^2}{E_i}$
Todos iguales	680	720	2,22
Todos distintos	31	10	44,10
Exactamente un par	289	270	1,34
	1000	1000	47,66

Observamos que $47.66 > \chi_{0.05;2}^2 = 5.99$ y por lo tanto rechazamos la hipótesis de independencia de los números.

Esta prueba se puede extender a más dígitos pero a su vez las posibilidades aumentan y los cálculos se complican. Por ejemplo, para 5 dígitos, podríamos tener, todos iguales, todos distintos, exactamente un par, exactamente un trio, un trio y un par, dos pares, etc.