Testowanie generatorów liczb losowych

- **Metodologia testowania generatorów liczb losowych**
- **Testy zgodności z rozkładem równomiernym.**
- **Testy zgodności rozkładów statystyk.**
- **Testy serii.**
- **Testy kombinatoryczne.**
- **Testowanie za pomocą zadań kontrolnych.**

$\Rightarrow$ http://th-www.if.uj.edu.pl/~placzek/dydaktyka/MMC/

"Random number generators should not be chosen at random."

Donald E. Knuth

Jak sprawdzić czy dany generator jest dobry?

▶ Generator jest dobry jeśli daje sekwencje liczb, które posiadają własności liczb prawdziwie losowych. ← Jak to sprawdzic?

● **Podejście tradycyjne:**

Sformułować pewne **własności** liczb losowych o rozkładzie równomiernym między 0 a 1, tzn. $r \in \mathcal{U}(0,1)$, i sprawdzić – przez wykonanie odpowiednich **testów** – czy sekwencje liczb z danego generatora posiadają te własności.

$\longrightarrow$ Ale można sformułować nieskończoną liczbę takich własności $\Rightarrow$ nieskończona liczba testów!

▷ W praktyce można jedynie udowodnić, że generator jest zły (nie spełnia pewnych testów), ale nie można udowodnić, że generator jest dobry (fakt, iż przeszedł pomyślnie n testów nie daje gwarancji, że przejdzie $(n+1)$-szy test, którym akurat może być nasz rowiązywany problem!).

▶ Testowanie generatorów $\longrightarrow$ selekcja negatywna:

Pomyślne przejście pewnej liczby testów tylko zwiększa nasze zaufanie do danego generatora, ale nie daje nam pewności co do jego zupełniej niezawodności!

Metodologia testowania generatorów liczb losowych

- **Typowy schemat testowania generatora liczb losowych o rozkładzie $\mathcal{U}(0,1)$:**

1. Startując z **losowo** wybranej liczby początkowej generujemy n kolejnych liczb (n – ustalone).
2. Obliczamy wartość pewnej statystyki testowej T.
3. Obliczamy $F(T)$, gdzie F jest dystrybuantą statystyki T, gdy hipoteza jest prawdziwa.
4. Powtarzamy powyższą operację N razy, obliczając w kolejnych krokach wartości statystyki: $T_1, T_2, \ldots, T_N$ oraz wartości: $F(T_1), F(T_2), \ldots, F(T_N)$.

Jeżeli weryfikowana hipoteza jest prawdziwa, to $F(T_1), F(T_2), \ldots, F(T_N)$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie $\mathcal{U}(0,1)$. Testowanie generatora kończymy testowaniem tej hipotezy (na określonym **poziomie istotności**).

▶ **Do dziś sformułowano wiele różnorodnych i surowych testów:**

▷ Patrz np. Donald E. Knuth, *„Sztuka programowania”*, tom 2, WNT 2002.

▷ Np. Bateria testów DIEHARD G. Marsaglii (http://www.stat.fsu.edu/pub/diehard/)
– pomogła wyeliminować wiele złych generatorów liczb losowych, w tym fizycznych.

- **Nowe (niestandardowe) podejście** – M. Lüscher (1993):

Użycie formalizmu stosowanego przy opisie chaotycznego zachowania w klasycznych układach dynamicznych do badania oraz konstrukcji generatorów liczb losowych (patrz poprzedni wykład).

▶ Generator **RANLUX** – spełnia wymagania „chaotyczności” według powyższego formalizmu.

▷ Spełnia też wszystkie znane testy statystyczne – nie bez powodu!

Test chi-kwadrat

Przedział $[0,1)$ dzielimy na k podprzedziałów: $0 = a_0 < a_1 < \ldots < a_k = 1$.

Niech n_i – liczba elementów ciągu $\{X_1, \ldots, X_n\}$ należących do przedziału $[a_{i-1}, a_i)$,

a $p_i \equiv P\{a_{i-1} \le X < a_i\} = a_i - a_{i-1}$, $(i = 1, \ldots, k)$ dla rozkładu $\mathcal{U}(0,1)$.

$\Rightarrow$ Zmienna losowa:

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad n = \sum_{i=1}^{k} n_i,$$

ma rozkład χ^2 (chi-kwadrat) o $(k-1)$ stopniach swobody.

▷ Wartość powyższej statystyki $\longrightarrow$ weryfikacja hipotezy, że: $X_1, \ldots, X_n \in \mathcal{U}(0,1)$.

▶ Uproszczenie: $p_i \equiv a_i - a_{i-1} = p_j \equiv a_j - a_{j-1} = \frac{1}{k}$, $i, j = 1, \ldots, k$ (równy podział):

$$\chi^2_{k-1} = \frac{k}{n} \sum_{i=1}^{k} n_i^2 \; - n \, .$$

¶ **Dygresja** – rozkład χ^2: $X \in \mathbb{R}$, $X > 0$, $N \in \mathbb{N}$, (N – liczba ,,stopni swobody"),

Gęstość prawdop.: $\rho(X) = \frac{1}{2}\left(\frac{X}{2}\right)^{\frac{N}{2}-1} e^{-\frac{X}{2}} \left[\Gamma\left(\frac{N}{2}\right)\right]^{-1}$; $E(X) = N$, $V(X) = 2N$.

Jeżeli $X_1, \ldots, X_N$ – niezależne zmienne o rozkładzie normalnym $N(0,1)$,

to $S_N = \sum_{i=1}^{N} X_i^2$ ma rozkład χ^2_N o N stopniach swobody.

Test zgodności z rozkładem wielowymiarowym

- Z kolejnych liczb otrzymywanych z generatora tworzymy wektory (punkty) m-wymiarowe:

$$(X_1, X_2, \ldots, X_m), (X_{m+1}, X_{m+2}, \ldots, X_{2m}), \ldots, (X_{(n-1)m+1}, X_{(n-1)m+2}, \ldots, X_{nm})$$

  $\longrightarrow$ powinny mieć rozkład równomierny na kostce $(0,1)^m$.

  Każdy przedział $(0,1)$ dzielimy na k równych podprzedziałów: $([j-1]/k, j/k)$, $j = 1, \ldots, k$.

  $\longrightarrow$ Tzn. kostkę dzielimy na k^m jednakowych kostek o objętościach k^{-m}.

  Niech n_i – liczba m-wymiarowych punktów, które wpadły do i-tej kostki $\Rightarrow$ **statystyka testu χ^2:**

$$\chi^2_{k^m-1} = \frac{k^m}{n} \sum_{i=1}^{k^m} n_i^2 - n\,, \qquad n = \sum_{i=1}^{k^m} n_i\,.$$

- Postępujemy jak wyżej, ale tworzymy m-wymiarowe punkty z nakładającymi się współrzędnymi:

$$(X_1, X_2, \ldots, X_m), (X_2, X_3, \ldots, X_{m+1}), (X_3, X_4, \ldots, X_{m+2}), \ldots.$$

  $\longrightarrow$ Dla N liczb losowych mamy $N - m + 1$ takich punktów.

  Definiujemy statystyki:

$$\psi_0^2 = 0, \qquad \psi_m^2 = \sum_{i=1}^{k^m} \frac{[n_i - (N-m+1)/k^m]^2}{(N-m+1)/k^m}, \quad m = 1, 2, \ldots.$$

  ▶ Dla ,,dostatecznie dużych" N zmienna losowa $(\psi_m^2 - \psi_{m-1}^2)$ ma w przybliżeniu rozkład chi-kwadrat o $(k^m - k^{m-1})$ stopniach swobody.

Test OPSO (ang. *overlaping-pairs-sparse-occupancy*)

▶ Test OPSO (G. Marsaglia, 1984) dotyczy analizy częstości nakładających się par liczb otrzymywanych z generatora.

Niech: $X_1, X_2, \ldots, X_n$ – ciąg liczb ciąg liczb z generatora. Z każdej liczby weźmy b bitów (np. najbardziej znaczących) i skonstruujmy ciąg liczb $I_1, I_2, \ldots, I_n$, gdzie $I_j \in \{0, 1, \ldots, 2^b - 1\}$.

Następnie utwórzmy ciąg kolejnych nakładających się par:

$$(I_1, I_2), (I_2, I_3), \ldots, (I_{n-1}, I_n).$$

Niech Y – liczba takich par ze zbioru $\{(i,j) : i,j = 0, \ldots, 2^b - 1\}$, które **nie pojawiły się** w powyższym ciągu.

▶ Zmienna losowa Y ma asymptotycznie ($n \to \infty$) rozkład normalny $N(\mu, \sigma)$.

Przykładowe parametry dla testu OPSO:

b	n	μ	σ
10	2^{21}	141 909	290.26
11	2^{22}	1 542 998	638.75
11	2^{23}	567 639	580.80

▷ Tego typu test można rozszerzyć na trójki (OTSO), czwórki (OQSO), itd. liczb z generatora (G. Marsaglia, 1993; patrz np. DIEHARD: http://www.stat.fsu.edu/pub/diehard/).

Test Kołmogorowa–Smirnowa (test K–S)

▶ Test K–S służy do weryfikacji hipotezy, że zmienna losowa X ma rozkład o danej ciągłej dystrybuancie F. Statystyka testu opiera się na różnicy między hipotetyczną dystrybuantą F a dystrybuantą empiryczną F_n z próby $X_1, X_2, \ldots, X_n$.

● Statystyka testowa:

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F(x)|, \qquad \text{gdzie: } F_n(x) = \frac{1}{n}\sum_{j=1}^{n} \Theta(x - X_j)\,.$$

▶ Jeżeli próba pochodzi z rozkładu o dystrybuancie F, to $D_n \to 0$ z prawdopodobieństwem 1.

$\longrightarrow$ Duże wartości statystyki D_n przemawiają przeciwko wyjściowej hipotezie!

▷ Wartości krytyczne testu $D_n(\alpha)$ dla danego poziomu istotności α, tzn.

$$\mathcal{P}_F\{D_n > D_n(\alpha)\} = \alpha$$

można znaleźć w tablicach statystycznych (nie zależą od postaci funkcji F).

▶ Dla rozkładu $\mathcal{U}(0,1)$ dystrybuanta wynosi:

$$F(x) = x, \quad 0 < x < 1.$$

Test K–S: obliczenia praktyczne

- **Spostrzeżenie:**

  Dystrybuanta empiryczna F_n jest funkcją schodkową i $\sup_{-\infty<x<+\infty} |F_n(x) - F(x)|$ osiągane jest w jednym z jej punktów skoku.

  $\Rightarrow$ Liczby $X_1, X_2, \ldots, X_n$ można posortować i wykonać obliczenia wg. wzorów:

$$D_n^+ = \max_{1\le i\le n} \left(\frac{i}{n} - F(X_{i:n}) \right), \qquad D_n^- = \max_{1\le i\le n} \left(F(X_{i:n}) - \frac{i-1}{n} \right),$$

$$D_n = \max\{D_n^+, D_n^-\}$$

  gdzie $X_{i:n}$ – statystyka pozycyjna, tzn. $X_{1:n} \le X_{2:n} \le \ldots \le X_{n:n}$.

- ▶ Statystyki D_n^+ i D_n^- mają identyczne rozkłady, które asymptotycznie ($n \gtrsim 80$) osiągają:

$$\lim_{n\to\infty} \mathcal{P}\{\sqrt{n} D_n^{\pm} \le t\} = 1 - e^{-2t^2}, \quad t > 0.$$

- ▶ Statystyka D_n asymptotycznie ($n \gtrsim 80$) osiąga **rozkład λ-Kołmogorowa**:

$$\lim_{n\to\infty} \mathcal{P}\{\sqrt{n} D_n \le t\} = K(t) \equiv \sum_{j=-\infty}^{\infty} (-1)^j \, e^{-2j^2 t^2}, \quad t > 0,$$

  którego wartości krytyczne λ_α ($\mathcal{P}\{\sqrt{n} D_n > \lambda_\alpha\} = \alpha$) można znaleźć w tablicach statystycznych.

  $\rightarrow$ Często używane wartości, to: $\lambda_{0.1} = 1.224,\ \lambda_{0.05} = 1.358,\ \lambda_{0.01} = 1.628$.

Ogólny schemat

Niech: $y = h(x_1, x_2, \ldots, x_m)$ – funkcja określona na kostce $(0,1)^m$, będąca zmienną losową, jeżeli jej argumenty są niezależnymi zmiennymi losowymi $\in \mathcal{U}(0,1)$.

▶ Z liczb otrzymywanych z generatora tworzymy ciąg:

$$Y_j = h(X_{(j-1)m+1}, X_{(j-1)m+2}, \ldots, X_{jm}), \quad j = 1, 2, \ldots$$

i weryfikujemy hipotezę, że jest on próbką prostą z populacji o dystrybuancie:

$$G(y) = \mathcal{P}\{Y_j \leq y\}.$$

Testy oparte na statystykach pozycyjnych

Bierzemy funkcje postaci:

$$u = \max\{x_1, \ldots, x_m\}, \qquad v = \min\{x_1, \ldots, x_m\}, \qquad r = u - v\,.$$

▶ Rozkłady odpowiednich zmiennych losowych U_j, V_j i R_j dane są:

$$\mathcal{P}\{U_j \leq u\} = u^m, \quad 0 \leq u \leq 1; \qquad \mathcal{P}\{V_j \leq v\} = 1 - (1-v)^m, \quad 0 \leq v \leq 1;$$

$$\mathcal{P}\{R_j \leq r\} = m r^{m-1} - (m-1) r^m, \quad 0 \leq r \leq 1.$$

▷ Weryfikujemy hipotezy o zgodności rozkładów zmiennych U_j, V_j i R_j, obliczanych w oparciu o generowane ciągi liczb $X_{(j-1)m+1}, \ldots, X_{jm}$ z powyższymi rozkładami teoretycznymi.

$\rightarrow$ Zwykle testy przeprowadza się dla kilku wartości $m = 2, 3, \ldots, 10$.

Test sum

- Funkcja h ma postać:

$$y = x_1 + x_2 + \ldots + x_m.$$

▶ Odpowiednie zmienne losowe Y_j mają rozkład o gęstości prawdopodobieństwa:

$$g_m(y) = \begin{cases} \frac{1}{m-1}\left[y^{m-1} - \binom{m}{1}(y-1)^{m-1} + \binom{m}{2}(y-2)^{m-1} - \ldots\right] & \text{dla } 0 \le y \le m, \\ 0 & \text{poza tym,} \end{cases}$$

gdzie sumowanie wykonuje się dopóki $y,\ y-1,\ y-2,\ldots$ są dodatnie.

▷ Dla $m = 2 \ \rightarrow$ rozkład trójkątny o gęstości:

$$g_2(y) = \begin{cases} y & \text{dla } 0 \le y \le 1, \\ 2 - y & \text{dla } 1 \le y \le 2. \end{cases}$$

▷ Dla $m = 3 \ \rightarrow$ rozkład o gęstości:

$$g_3(y) = \begin{cases} \frac{1}{2}y^2 & \text{dla } 0 \le y \le 1, \\ \frac{1}{2}\left[y^2 - 3(y-1)^2\right] & \text{dla } 1 \le y \le 2, \\ \frac{1}{2}\left[y^2 - 3(y-1)^2 + 3(y-2)^2\right] & \text{dla } 2 \le y \le 3. \end{cases}$$

▶ Dla dużych m rozkład sum przybliża się rozkładem normalnym (na podstawie CTG).

Test d^2

- Dla $m = 4$ definiujemy funkcję h w postaci:

$$y = (x_1 - x_3)^2 + (x_2 - x_4)^2,$$

tzn. jest to kwadrat odległości między punktami (x_1, x_2) i (x_3, x_4) w kwadracie $(0, 1)^2$.

▶ Jeżeli niezależne zmienne losowe $X_1,\ X_2,\ X_3, X_4 \in \mathcal{U}(0, 1)$, to zmienna losowa

$$d^2 = (X_1 - X_3)^2 + (X_2 - X_4)^2$$

ma rozkład dany wzorem:

$$\mathcal{P}\{d^2 \le y\} = \begin{cases} \pi\, y - \frac{8}{3}\, y^{\frac{3}{2}} + \frac{1}{2}\, y^2 & \text{dla } 0 \le y \le 1, \\ \frac{1}{3} + (\pi - 2)y + 4(y-1)^{\frac{1}{2}} + \frac{8}{3}(y-1)^{\frac{3}{2}} & \\ -\frac{1}{2}\, y^2 - 4y \operatorname{arc\,sec}(y^{\frac{1}{2}}) & \text{dla } 1 \le y \le 2. \end{cases}$$

▷ Testowanie generatora polega na weryfikacji zgodności rozkładu statystyki d^2 obliczanej dla liczb przez niego generowanych z powyższym rozkładem teoretycznym.

Test urodzin odstępów

- Niech ciąg liczb pseudolosowych z generatora: $I_1, I_2, \ldots, I_m \in \{1, 2, \ldots, n\}$.

  $\longrightarrow$ Sortujemy powyższy ciąg w ciąg niemalejący: $I_{1:m}, I_{2:m}, \ldots, I_{m:m}$,

  a następnie tworzymy ciąg **odstępów**:

$$I_{1:m},\ I_{2:m} - I_{1:m}\ , I_{3:m} - I_{2:m},\ \ldots,\ I_{m:m} - I_{m-1:m}\ .$$

- Niech: Y – liczba odstępów, które występują więcej niż raz w powyższym ciągu.

  ▶ Jeżeli $I_1, \ldots, I_m \in \{1, 2, \ldots, n\}$ – niezależne zmienne losowe o rozkładzie równomiernym na zbiorze $\{1, 2, \ldots, n\}$, to zmienna losowa Y ma **rozkład Poissona** z parametrem:

$$\mu = \frac{m^3}{4n}\ .$$

  ▷ W praktyce bierze się $n \geq 10^4$, a wartości I_j tworzy się z najbardziej znaczących bitów liczb z generatora.

$\longrightarrow$ Tego testu nie spełniają zwykle generatory Fibonacciego typu: $F(r, s, \pm)$ i $F(r, s, \mathrm{xor})$!

¶ **Dygresja** – rozkład Poissona $P(\mu)$:

$$\mathcal{P}[X = k] = \frac{\mu^k}{k!}\, e^{-\mu}\ , \quad k = 0, 1, \ldots; \qquad E(k) = V(k) = \mu\ .$$

Test najmniejszej odległości w parach

Generujemy n punktów z kostki $(0,1)^m$. Bierzemy $\binom{n}{2}$ par punktów i obliczamy odległość Euklidesową między punktami każdej pary.

Niech D – najmniejsza odległość między parami punktów $\longrightarrow$ dla równomiernego rozkładu liczb z generatora zmienna losowa: $\boxed{T = n^2 D^m/2}$ ma asymtotycznie rozkład wykładniczy ze średnią: $1/V_m$, gdzie V_m – objętość m-wymiarowej kuli jednostkowej.

Schemat testu:

- Generujemy Nn punktów w kostce $(0,1)^m$, uzyskując N realizacji statystyki T.
- Porównujemy rozkład empiryczny T z rozkładem wykładniczym, np. stosując test K–S lub test chi-kwadrat.

▷ Uwaga: Aby jakość testu była dobra, wartości N, n i m powinny być odpowiednio dobrane!

▶ Przykładowe wartości parametrów:

1) $N = 100,\ n = 10^5,\ m = 4;$
2) $N = 20,\ n = 10^5,\ m = 6;$
3) $N = 20,\ n = 5 \cdot 10^4,\ m = 9.$

▷ Generatory liniowe zwykle nie spełniają tego testu, gdyż tworzą regularne siatki w $\mathbb{R}^m$.

¶ **Dygresja** – rozkład wykładniczy $E(\lambda)$: $\rho_\lambda(X) = \lambda\, e^{-\lambda X},\ x \geq 0, \quad E(X) = \sigma(X) = 1/\lambda\,.$

Testy serii

Niech X – zmienna losowa o rozkładzie danym dystrybuantą F. Zbiór wartości tej zmiennej dzielimy na dwa rozłączne podzbiory A i B i definiujemy zmienną losową Y o wartościach a i b:

$$Y = \begin{cases} a & \text{gdy } X \in A, \\ b & \text{gdy } X \in B. \end{cases}$$

Ciąg liczb losowych $X_1, X_2, \ldots, X_N$ przekształcamy w ciąg $Y_1, Y_2, \ldots, Y_N \in \{a, b\}$.

▶ **Seria** – każdy odcinek ciągu złożony z maksymalnej liczby kolejnych jednakowych elementów.

▷ Np. dla ciągu: $a, a, b, a, b, b, b, a, b, b, b$ mamy następujące serie: $aa,\ b,\ a,\ bbb,\ a,\ bbb$.

Niech n_a – liczba symboli a w danym ciągu $Y_1, Y_2, \ldots, Y_N$, a $n_b = N - n_a$ – liczba symboli b.

$\Rightarrow$ Rozkład liczby serii R przy tym warunku dany jest wzorem:

$$\mathcal{P}\{R = r | n_a, n_b\} = \begin{cases} 2\binom{n_a-1}{k-1}\binom{n_b-1}{k-1} \Big/ \binom{N}{n_a} & \text{gdy } r = 2k, \\ \left[\binom{n_a-1}{k}\binom{n_b-1}{k-1} + \binom{n_a-1}{k-1}\binom{n_b-1}{k}\right] \Big/ \binom{N}{n_a} & \text{gdy } r = 2k+1. \end{cases}$$

$\Rightarrow$ Rozkład bezwarunkowy liczby serii R:

$$\mathcal{P}\{R = r\} = \sum_{n_a=0}^{n_a+n_b} \mathcal{P}\{R = r | n_a, n_b\} \binom{n_a + n_b}{n_a} p^{n_a} (1-p)^{n_b}, \qquad p \equiv \mathcal{P}\{Y = a\}.$$

Obliczenia praktyczne

Weryfikacja hipotezy o niezależości zmiennych $X_1, X_2, \ldots, X_N$:

- Dla ustalonego poziomu istotności α należy znaleźć dwie wartości krytyczne R_1 i R_2:

$$\mathcal{P}\{R < R_1\} = \mathcal{P}\{R > R_2\} = \frac{\alpha}{2}.$$

- Liczby R_1 i R_2 otrzymujemy rozwiązując równania:

$$\sum_{j=0}^{R_1-1} \mathcal{P}\{R = j\} = \frac{\alpha}{2}, \qquad \sum_{j=R_2+1}^{N} \mathcal{P}\{R = j\} = \frac{\alpha}{2}$$

  ▷ Rozwiązania tych równań można znaleźć w tablicach statystycznych.

- Dla dużych N rozkład liczby serii aproksymuje się rozkładem normalnym $N(\mu, \sigma)$:

$$\mu = E(R) = 2Np(1-p) + p^2 + (1-p)^2,$$

$$\sigma^2 = V(R) = 4Np(1-p)[1-3p(1-p)] - 2p(1-p)[3-10p(1-p)].$$

- Jeżeli zaobserwowana liczba serii $R < R_1$ lub $R > R_2$, to weryfikowaną hipotezę odrzucamy.

▶ Test **serii względem mediany**, tzn. $A = (0, 0.5),\ B = (0.5, 1),\ p = 1/2$:

→ Dla dużych N rozkład normalny z parametrami: $\mu = N/2,\ \sigma^2 = N/4.$

▶ Test **serii monotonicznych**, tzn. ciągów znaków: $\mathrm{sign}(X_2 - X_1), \mathrm{sign}(X_3 - X_2), \ldots \in \{+, -\}$

→ Dla dużych N rozkład normalny z parametrami: $\mu = (2N-1)/3,\ \sigma^2 = (16N-29)/90.$

▶ Testy kombinatoryczne należą do grupy tzw. **testów niezależności (losowości próby)**.

Test pokerowy

- Przedział zmiennych losowych X dzielimy na k jednakowych podprzedziałów:
  $0 = a_0 < a_1 < \ldots < a_k = 1$.
  ▷ Dla ciągu liczb losowych $X_1, X_2, \ldots, X_n \in \mathcal{U}(0,1)$:
  $$\mathcal{P}\{a_{i-1} < X_j \leq a_i\} = \frac{1}{k}.$$
- Tworzymy ciąg zmiennych losowych Y_j według wzoru:
  $$Y_j = i, \text{ jeżeli } X_j \in (a_i, a_{i+1}), \quad i = 0, 1, \ldots, k-1.$$
  ▷ Zmienna losowa Y_j przyjmuje każdą wartość z jednakowym prawdopodobieństwem.
- Ciąg $Y_1, Y_2, \ldots$ dzielimy na piątki:
  $$(Y_1, Y_2, \ldots, Y_5), (Y_6, Y_7, \ldots, Y_{10}), \ldots .$$
  ▷ Nowy ciąg jest zbudowany z k^5 różnych piątek.
- Wyróżniamy następujące typy piątek:
  $abcde$ (bust), $aabcd$ (para), $aabbc$ (dwie pary), $aaabc$ (trójka),
  $aaabb$ (full), $aaaab$ (czwórka), $aaaaa$ (piątka).

Test pokerowy – rozkłady

Jeżeli $X_1, X_2, \ldots, X_n \in \mathcal{U}(0,1)$ – niezależne zmienne losowe, to rozkłady poszczególnych typów piątek dane są wzorami:

$$\mathcal{P}\{(abcde)\} = \frac{(k-1)(k-2)(k-3)(k-4)}{k^4}, \quad k \geq 5,$$

$$\mathcal{P}\{(aabcd)\} = \frac{10(k-1)(k-2)(k-3)}{k^4}, \quad k \geq 4,$$

$$\mathcal{P}\{(aabbc)\} = \frac{15(k-1)(k-2)}{k^4}, \quad k \geq 3,$$

$$\mathcal{P}\{(aaabc)\} = \frac{10(k-1)(k-2)}{k^4}, \quad k \geq 3,$$

$$\mathcal{P}\{(aaabb)\} = \frac{10(k-1)}{k^4}, \quad k \geq 3,$$

$$\mathcal{P}\{(aaaab)\} = \frac{5(k-1)}{k^4}, \quad k \geq 2,$$

$$\mathcal{P}\{(aaaaa)\} = \frac{1}{k^4}, \quad k \geq 1,$$

▷ W praktyce najczęściej używa się wartości $k = 2, 8, 10$.

▶ Zgodność rozkładu piątek różnych typów sprawdza się przy pomocy testu chi-kwadrat.

Test kolekcjonera

- Tworzymy ciąg $Y_1, Y_2, \ldots,$ jak dla testu pokerowego.
- Obserwujemy ten ciąg aż pojawią się w nim wszystkie k liczby: $0, 1, \ldots, k-1$.

▶ Długość zaobserwowanego odcinka ciągu R ma rozkład:

$$\mathcal{P}\{R = r\} = \frac{1}{k^{r-1}} \sum_{j=0}^{k-2} (-1)^j \binom{k-1}{j} (k-1-j)^{r-1}, \qquad r = k, k+1, \ldots$$

▷ Zgodność rozkładu zaobserwowanego w powyższym rozkładem teoretycznym weryfikuje się przy użyciu standardowego testu chi-kwadrat.

Test permutacji

- Bierzemy ciąg k-wymiarowych punktów utworzony z nk kolejnych liczb z generatora:

$$(X_1, X_2, \ldots, X_k), (X_{k+1}, \ldots, X_{2k}), \ldots, (X_{k(n-1)+1}, , \ldots, X_{nk})$$

i każdy z punktów przekształcamy zastępując współrzędne ich **rangami** (tzn. numerami porządkowymi w kolejności rosnącej wartości współrzędnych).

▶ Każda permutacja $(n_1, n_2, \ldots, n_k)$ liczb $(1, 2, \ldots, k)$ jest jednakowo prawdopodobna.

▷ Powyższą hipotezę weryfikuje się standardowym testem chi-kwadrat.

Test kolizji

- Bierzemy ciąg utworzony z nk kolejnych liczb z generatora:
$$(X_1, X_2, \ldots, X_k), (X_{k+1}, \ldots, X_{2k}), \ldots, (X_{k(n-1)+1}, , \ldots, X_{nk}).$$
- Kostkę $(0,1)^k$ dzielimy na $m = s^k$ jednakowych k-wymiarowych kostek o objętościach $1/m$.
- Za statystykę testową przyjmujemy **liczbę kolizji** C, tzn. liczbę przypadków, w których kolejny punkt wpada do kostki zajętej już przez co najmniej jeden punkt.
- Jeżeli $X_1, X_2, \ldots, X_n \in \mathcal{U}(0,1)$ – niezależne zmienne losowe, to statystyka C ma rozkład:
$$\mathcal{P}\{C = c\} = \frac{m(m-1)\ldots(m-n-c+1)}{m^n} \binom{n}{n-c}, \qquad c = 0, 1, \ldots, n.$$

▷ Do weryfikacji zgodności rozkładu zaobserwowanego w powyższym rozkładem teoretycznym używa się testu chi-kwadrat.

Test pustych komórek (test Davida–Hellwiga)

▶ Statystyka C' – liczba pustych komórek, tzn. kostek, w których nie ma ani jednego punktu.

$\Rightarrow$ Rozkład prawdopodobieństwa:
$$\mathcal{P}\{C' = c\} = \binom{m}{c} \sum_{i=0}^{m-c} (-1)^i \binom{m-c}{i} \left[1 - \frac{c+i}{m}\right]^m, \qquad c = 0, 1, \ldots, m.$$

▷ Wartości dystrybuanty oraz wartości krytyczne można znaleźć w tablicach statystycznych.

Testowanie za pomocą zadań kontrolnych

▶ Testowanie generatorów za pomocą **zadań kontrolnych** – rozwiązywanie określonych zadań metodami Monte Carlo i porównywanie wyników z wynikami otrzymywanymi w inny sposób.

- Wyznaczanie liczby π, np. metodą igły Buffona lub trafiania w ćwiartkę jednostkowego koła zawartą w jednostkowym kwadracie.
- Obliczanie znanych całek metodami Monte Carlo.
- Obliczanie objętości kuli jednostkowej w przestrzeni m-wymiarowej:

$$V_m = \frac{2\pi^{\frac{m}{2}}}{m\,\Gamma\left(\frac{m}{2}\right)}, \quad \Gamma(x+1) = x\Gamma(x),\ \ \Gamma(1) = 1,\ \ \Gamma\left(1/2\right) = \sqrt{\pi}$$ (funkcja Γ-Eulera).

  ▷ Można zastosować metodę ,,orzeł–reszka” generując punkty losowe w kostce $(-1, 1)^m$.

- Wyznaczanie parametrów pewnych zjawisk fizycznych, których ścisłe rozwiązania są znane (np. ruchy Browna na płaszczyźnie).

⇛ Ćw. N7.1: Wykonać następujące testy zaimplementowanych generatorów liczb losowych:

(a) Test chi-kwadrat, np. dla $k = 10$.

(b) Jeden z testów zgodności z rozkładem wielowymiarowym, np. dla $m = 3$.

(c) Test Kołmogorowa–Smirnowa.

(d) Po jednym z testów: zgodności rozkładów statystyk, serii oraz kombinatorycznych.

(e) Objętości kuli m-wymiarowej dla kilku wartości m (wyliczać też odchylenie standardowe).

⇛ Ćw. N3* (nadobowiązkowe): Zaimplementować i wykonać test OPSO.