Arbitrary p.d.f. **generation**

Marcin Chrząszcz
mchrzasz@cern.ch

University of Zurich[UZH]

Monte Carlo methods,
7 April, 2016

Reverting the c.d.f.

⇛ Let U be a random variable from $\mathcal{U}(0,1)$
⇛ Now let F be a non decreasing function such that:

$$F(-\infty) = 0 \quad F(\infty) = 1$$

then:

$$X = F^{-1}(U)$$

has a p.d.f. distribution with a c.d.f. function of F.

⇛ Prove:

$$F(x) = \mathcal{P}(U \leqslant F(x)) = \mathcal{P}(F^{-1}(U) \leqslant x) = \mathcal{P}(X \leqslant x) \quad \square$$

⇛ So it looks very simple if $x_1, X_2, ..., X_n$ are random variables from $\mathcal{U}(0,1)$ then: $\{X_i = F^{-1}(x_i)\}, i = 1, ..., n$ is the sequence that has a c.d.f. distirbution of F.

Reverting the c.d.f., examples

- ⇛ The exponential distribution: $E(0,1)$.
- ⇛ The p.d.f.: $\rho(X) = e^{-X}, \ X \geqslant 0$.
- ⇛ The c.d.f.: $F(x) = \int_0^x e^{-X} dX = 1 - e^{-x}$.
- ⇛ Now let $R \in \mathcal{U}(0,1): \ R = F(X) = 1 - e^{-X} \longrightarrow X = -\ln(1-R)$
- ⇛ Now we can play a trick: if $R \in \mathcal{U}(0,1)$ then $1-R$ also in $\mathcal{U}(0,1)$.
- ⇛ In the end we get: $X = -\ln(R)$

- ⇛ Use the reverting to generate the following distributions:

- E 6.1 $\rho(X) = \dfrac{c}{X}$ on the interval $[a,b]$, where $0 < a < b < \infty$.
- E6.2 The Breit-Wigner function:

$$\rho_{\theta,\lambda}(X) = \frac{\lambda}{\pi}\frac{1}{(X-\theta)^2+\lambda^2}$$

Hit: First do $C(0,1)$ then transform the variables.

Reverting the c.d.f., general case

The χ^2 texts with $\mathcal{U}(0,1)$

⇛ The algorithm:

- Divide the $[0,1)$ into k subdivisions:

$$0 = a_0 < a_1 < a_2 < a_3 < ... < a_k = 1$$

- Let $a_{n_i} = X_1, X_2, ..X_n$ be an series of elements in the interval $[a_{i-1}, a_i)$ (with n_i elements). The $p_i = Pa_{i-1} < X < a_i = a_i - a_{i-1}$.
- A random variable:

$$\chi^2_k = \sum_{i=1}^{k} \frac{(n_1 - np_i)^2}{np_i}, \quad n = \sum_{i=1}^{k} n_i,$$

had a χ^2 distribution of $k-1$ degrees of freedom.

⇛ The above hypothesis verifies if the random numbers are indeed $\mathcal{U}(0,1)$.

⇛ The χ^2 distribution: $X \in \mathbb{R}, X > 0, N \in \mathbb{N}$:

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

The multi dimension test

⇛ From the obtained numbers we construct an m dimension points:

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

- In principle they should have a uniform distribution in an $(0,1)^m$ hipercube.
- we divide each edge of the hipercube into k equal subdivisions: $[j-1]/k, j/k),\ j = 1...k$.
- Now: n_i is the number of m dimensional points, which are in the i-th hipercube.
- The χ^2 test statistics:

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

⇛ Now we construct other points:

$$(X_1, X_2, ..., X_m), (X_2, X_3, ..., X_{m+1}), ...$$

- For N random numbers we have $N-m+1$ such numbers.
- We define the statistics:

$$\psi_0^2 = 0, \quad \psi_m^2 = \sum_{i=1}^{k^m} \frac{[n_1 - (N-m+1)/k^m]^2}{(N-m+1)\,k^m}, \quad m = 1, 2, ...$$

- For large N the $(\psi_m^2 - \psi_{m-1}^2)$ has a χ^2 distribution with $k^m - k^{m-1}$ degrees of freedom.

Overlaping-pairs-sparse-occupancy

⇛ The OPSO (G.Marsaglia 1984)is an analysis of pairs obtained from random number generator.

$X_1, X_2, ..., X_n$ - n random numbers obtained from generator. From each number we take b bits from which we construct a second series: $I_1, I_2, ..., I_n$, where $I_j \in \left[0, 1, ..., 2^b - 1\right]$.

⇛ Next we create the pair series:

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

⇛ Y - number of pairs from : $(i, j) : i, j = 0, 1, ..., 2^b - 1$, which DIDN'T occur in the above series.

Bitstream	No. missing words	z-score	p-value
23 to 32	141989	0.2747	0.391764
22 to 31	142538	2.1678	0.015086
21 to 30	142084	0.6023	0.273484
20 to 29	142081	0.5920	0.276937

⇛ This kind of test can be exteded to triple-pairs, and quadro-pairs.
⇛ See DIEHARD G.Marsgalia 1993 http://stat.fsu.edu/pub/diehard/

Kolomogorox-Smirnow

$\Rrightarrow$ The K-S test is used to check if a Random variable has pdf of a distribution F. The test is based on the difference between the two distributions:

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

$\Rrightarrow$ If the random generator is from the F distribution then the $D_n \rightarrow 0$ with the probability 1.
$\Rrightarrow$ Large values of D_n exclude the generator. $\Rrightarrow$ The critical values of the test $D_n(\alpha)$ can be find in the mathematical tables for every α:

$$\mathcal{P}[D_n < D_n(\alpha)] = \alpha$$

$\Rrightarrow$ They do not depend on the F function.
$\Rrightarrow$ For the $\mathcal{U}(0,1)$:

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

Kolomogorox-Smirnow in practice

Take note:

Empirical CDF of F_n is a step function and $\sup_{-\infty<x<\infty} |F_n(x) - F(x)|$ is achieved only in one point!

⇛ In practice one should sort the numbers: $X_1, ..., X_n$ and calculate the following:

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

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

where $X_{i:n}$ is so-called position statistic: $X_{1:n}, X_{2:n}, ..., X_{i:n}$.

⇛ The statistic D_n asymptotically (in practice $n \geqslant 80$) is approaching the λ-Kolomogorows cdf:

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

for which the critical values $\lambda_\alpha(\mathcal{P}\{\sqrt{n} D_n > \lambda_\alpha$ can be found in the mathematical tables.

⇛ Commonly the $\lambda_{0.1} = 1.224$, $\lambda_{0.05} = 1.358$, $\lambda_{0.01} = 1.628$ are used.

Statistic distributions test- sum test

⇛ The h function has the form:

$$y = x_1 + x_2 + x_3 ... x_m.$$

⇛ the random variables form the new pdf:

$$g_m(y) = \begin{cases} \frac{1}{m-1}\left[y^{m-1} - \binom{m}{1}(y-1)^{m-1} + \binom{m}{2}(y-2)^{m-1} - ..\right] & \text{for} 0 \leqslant y \leqslant m, \\ 0 & \text{else} \end{cases}$$

where you stop when $y - m$ is negative. ⇛ For $m = 2$ we have the triangle pdf:

$$g_2(y) = \begin{cases} y, \ \text{for } 0 \leqslant y \leqslant 1 \\ 2 - y, \ \text{for } 0 \leqslant y \leqslant 1 \end{cases}$$

⇛ For $m = 3$ we have the triangle pdf:

$$g_3(y) = \begin{cases} \frac{1}{2}y^2, \ \text{for } 0 \leqslant y \leqslant 1 \\ \frac{1}{2}\left[y^2 - 3(y-1)^2\right], \ \text{for } 1 \leqslant y \leqslant 2 \\ \frac{1}{2}\left[y^2 - 3(y-1)^2 3(y-2)^2\right], \ \text{for } 2 \leqslant y \leqslant 3 \end{cases}$$

⇛ For large m the g_m approaches the normal distribution.

Statistic distributions test- d^2

$\Rrightarrow$ for $m = 4$ we define the h:

$$y = (x_1 - x_3)^2 + (x_2 - X_4)^2$$

aka the square distance between (x_1, x_2) and (x_3, x_4).
$\Rrightarrow$ If the X_1, X_2, X_3, X_4 are from $\mathcal{U}(0,1)$ then:

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

had a pdf given by the following formula:

$$\mathcal{P}(d^2 - y) = \begin{cases} \pi y - \frac{8}{3} y^{\frac{3}{2}} + \frac{1}{2} y^2 & \text{for } 0 \leqslant y \leqslant 1 \\ -\frac{1}{2} y^2 - 4\mathrm{arcsec}(y^{\frac{1}{2}}) & \text{for } 1 \leqslant y \leqslant 2 \end{cases}$$

$\Rrightarrow$ Test is to check if the generated numbers have the aforementioned distribution.

Statistic distributions test- pair distance

$\Rrightarrow$ Generate n points from $(0,1)^m$. We take $\binom{n}{2}$ pairs of points and we calculate the distance between them.
$\Rrightarrow$ If D is the smallest distance between the pairs $\longmapsto$ for the $\mathcal{U}(0,1)^m$ the $T = n^2 D^m / 2$ has the exponential distribution with the mean $1/V_m$, where V_m is the hiper volume of the unite ball.
$\Rrightarrow$ In Patrice:

- We generate Nn points in the hipercube $(0,1)^m$, getting N points in the T statistics.
- We compare the empirical distribution T with the exponential distribution.
- WARNING: the N, n, m need to be choose smartly for the test to make sense.

$\Rrightarrow$ Linear generators usually fail this test!

Statistic distributions test- series test

$\Rrightarrow$ Lets assume our numbers are generated with a CDF F. The values of F we divide in two separated sub-samples: A and B.
$\Rrightarrow$ Furthermore we define the new variables Y such as:

$$Y = \begin{cases} = aX & \in A \\ = bX & \in B \end{cases}$$

$\Rrightarrow$ The random number sequence we transform the $X_1, X_2, X_3, ..., X_n$ into $Y_1, Y_2, Y_3, ..., Y_n$.
$\Rrightarrow$ Next we make series: For example the $a, a, b, a, a, b, b, b, a$ will be grouped into aa, b, aa, bbb, a.
$\Rrightarrow$ Let n_a be number of a symbols in $Y_1, Y_2, Y_3, ..., Y_n$. $n_b = N - n_a$.
$\Rrightarrow$ Distribution of number of series (R) is given by the equation:

$$\mathcal{P}(R = r, n_a, n_b) = \begin{cases} 2\binom{n_a-1}{k-1}\binom{n_b-1}{k-1}/\binom{N}{n_a} \text{ if } r = 2k \\ [\binom{n_a-1}{k}\binom{n_b-1}{k-1} + \binom{n_a-1}{k-1}\binom{n_b-1}{k}]/\binom{N}{n_a} \text{ if } r = 2k+1 \end{cases}$$

Statistic distributions test- poker test

⇛ The values of X random variable we divide into k identical sub samples:

$$0 < a_1 < ... < a_k = 1$$

⇛ For $X_1, X_2, ..., X_n$ from $\mathcal{U}(0,1)$:

$$\mathcal{P}(a_{i-1} < X_j < a_i) = \frac{1}{k}.$$

⇛ We create the new variables Y_1 accordingly:

$$Y_j = i \text{ if } X_j \in (a_{i-1}, a_i),\ i = 0, 1, ...k - 1$$

⇛ Now we create "the fives":

$$(Y_1, Y_2, Y_3, Y_4, Y_5), (Y_6, ...$$

⇛ There are couple of types of fives:

aabcd pair

aaabc three

aaaab four

aaaaa five

Statistic distributions test- poker test

⇛ If the variables are independent then we can calculate the probability:

$$\mathcal{P}\{(abcde)\} = \frac{(k-1)(k-2)(k-3)(k-4)}{k^4}, \quad k \geq 5,$$
$$\mathcal{P}\{(aabcd)\} = \frac{10(k-1)(k-2)(k-3)}{k^4}, \quad k \geq 4,$$
$$\mathcal{P}\{(aabbc)\} = \frac{15(k-1)(k-2)}{k^4}, \quad k \geq 3,$$
$$\mathcal{P}\{(aaabc)\} = \frac{10(k-1)(k-2)}{k^4}, \quad k \geq 3,$$
$$\mathcal{P}\{(aaabb)\} = \frac{10(k-1)}{k^4}, \quad k \geq 3,$$
$$\mathcal{P}\{(aaaab)\} = \frac{5(k-1)}{k^4}, \quad k \geq 2,$$
$$\mathcal{P}\{(aaaaa)\} = \frac{1}{k^4}, \quad k \geq 1,$$

⇛ In practice people choose: $k = 2, 8, 10$

⇛ The agreemnt of the distribution of different types of fives is check using the χ^2 test.

Conclusions

⇛ There are infinite number of tests one can invent for the testing of the generators.
⇛ All of the tests are in the same taste: invent a problem where you know the analytic solution, solve the problem and compare the results.
⇛ Homework: Use one of the previously implemented random number generator and :

- E5.1 Test them with chi-square test k=10.
- E5.2 Kolomorov-smirnon.
- E5.3 Multidimensional test.

Backup