Statystyka 1 – elementarne pojęcia

22.X.2014r.

Doświadczeniem losowym nazwiemy taki eksperyment, którego wyniku nie jesteśmy w stanie jednoznacznie z góry określić. W konsekwencji doświadczenia losowego następuje *elementarne zdarzenie losowe* – jest ono poprzez doświadczenie wybierane ze zbioru wszystkich możliwych zdarzeń; poszczególne elementarne zdarzenia losowe są rozłączne, wykluczają się wzajemnie.

Określmy na zbiorze zdarzeń elementarnych funkcję, odwzorowującą zdarzenia w liczby rzeczywiste.

Definicja (D1). Funkcję $\Omega \rightarrow \mathbb{R}$ nazywamy *zmienną losową*.

Przykład (P1). Niech $\Omega := \{\omega_1; \omega_2; \omega_3; \omega_4; \omega_5\}$, gdzie $X(\omega)$ dla kolejnych elementów ω_i jest równe, odpowiednio, 1, -2, 1, 3, 4.

Statystyka jako dział zajmuje się zmiennymi losowymi oraz ich funkcjami rzeczywistymi (funkcja zmiennej losowej jest funkcją losową).

Zmienne losowe opisują wybraną, pojedynczą cechę pewnych przedmiotów albo osób, np. wzrost człowieka, kolor samochodu, zarejestrowaną energię cząstki. Jeśli mamy do czynienia ze wszystkimi przedmiotami lub osobami o danej cesze bez wyjątku (z *populacją),* to funkcje zmiennych losowych na zbiorze cech populacji nazwiemy *parametrami* populacji. Jeśli zaś – tak jak jest w rzeczywistości – w doświadczeniu losowym dysponujemy możliwością zmierzenia interesującej nas cechy tylko u części (znikomej) populacji, zwanej *próbą*, funkcje zmiennych losowych określimy mianem *statystyk* próby.

Zasadniczym celem istnienia statystyki jest orzekanie o właściwościach populacji (np. wszystkich elektronów we Wszechświecie), czyli o jej parametrach, mając do dyspozycji jedynie statystyki pochodzące z próby (np. doświadczenia, w którym zarejestrowano tysiąc elektronów). Intuicyjnie oczywisty jest fakt, że im liczniejsza próba, tym statystyki są bliższe parametrom, a więc bardziej „wiarygodne”, lepiej orzekające o stanie faktycznym danej właściwości populacji.

W kolejnym kroku wprowadza się fundamentalne pojęcie *prawdopodobieństwa (pp),* w sposób aksjomatyczny. Mówi się o pp wylosowania w wyniku doświadczenia, elementu pewnego podzbioru A zdarzeń elementarnych, p(A). Jest to pojęcie, z którym Czytelnik wielokrotnie się już stykał w szkole, a zarazem tak obszerne w treści i znaczeniu, że zostanie tu uznane za dobrze znane i rozumiane.

W konsekwencji, następuje fundamentalna dla nauki statystyki

Definicja (D2). Dla zmiennej losowej X wprowadźmy funkcję, zwaną *dystrybuantą* tej zmiennej, określoną wzorem: $F_X(x) := p(A),\ A:=\{\omega:\ X(\omega)<x\}$.

Ćwiczenie (C1), czas: 25 minut. a) W przykładzie P1 podać $F_X(3)$, zakładając, że znamy wartości $p(\omega_i)$.

b) Niech $\bar{\bar{\Omega}}=2$. $X(\omega_1) = 1$, $X(\omega_2) = 2$. $p(\omega_1) = p(\omega_2) = ½$. Naszkicować wykres $F_X(x)$.

c) Naszkicować wykres $F_X(x)$ dla kostki sześciennej, tj. gdzie wartość X odpowiada liczbie wyrzuconych oczek dla każdego z sześciu zdarzeń elementarnych.

d) Naszkicować wykres $F_X(x)$ dla przykładu P1, jeśli każde ze zdarzeń ω_i jest jednakowo pp.

Stwierdzenia (W1). Dla upraszczającego zapis przypadku rozszerzonej prostej $\hat{\mathbb{R}}$, zawierającej nieskończoności: a) $F(-\infty) = 0$.

b) $F(\infty) = 1$.

c) $x \le y \Leftrightarrow F_X(x) \le F_X(y)$.

d) $p(X\in(a;b)) = F_X(b)-F_X(a)$.

Ćwiczenie (C2), czas: 10 minut. Udowodnić d), wychodząc z równości $1 = p(\Omega) = p(\ X\in(-\infty;+\infty)\ ) = \ldots$

Drugą podstawową funkcją o fundamentalnym znaczeniu dla fizyki, niosącą w sobie tę samą informację, co F_X oraz X, jest *funkcja rozkładu (funkcja gęstości pp)* zmiennej X:

Definicja (D3). Funkcję $f_X(x)$ taką, że $F_X(x) = \int_{-\infty}^{x} f_X(\xi)\,d\xi$, nazywamy *funkcja rozkładu (funkcja gęstości pp)* zmiennej X.

Można powiedzieć, że $f_X(x) \equiv \frac{dF_X}{dx}$.

Fundamentalną cechą $f_X(x)$ jest mierzenie pp: $p(X\in[a;b]) = \int_a^b f_X(x)dx$ (oczywisty dowód z D3 i D2, Czytelnik go wykona); jest ona zatem w dosłownym tego słowa znaczeniu gęstością pp. Argument *x* to ciągła wartość zmiennej losowej X, a $f_X(x)$ - gęstość pp., że zmienna X ma wartość dokładnie równą *x*. (Gęstość pp, a nie samo pp, bowiem pp trafienia w pojedynczy punkt na odcinku jest równe 0 – punkt jest nieskończenie krótszy od odcinka!)

W wypadku funkcji zmiennych dyskretnych, a nie ciągłych, wystarczy skorzystać z delty Diraca, np. dla kostki sześciennej: $f(x) = \frac{1}{6}\sum_{i=1}^{6} \delta(x-i)$.

Gwoli przypomnienia (lub informacji) Czytelnika: delta Diraca jest taką dystrybucją, że $\int_{-\infty}^{+\infty} y(x)\ \delta(a)\ dx = y(a)$; przy okazji $\int_{A \not\ni a} \delta(a)\ dx = 0$.

Ćwiczenie (C3), czas: 15 minut. Niech dystrybuanta zmiennej losowej X będzie funkcją ciągłą: x^2 na przedziale [0; 1] i niech wynosi 0 dla x < 0, a 1 dla x > 1.

a) naszkicować wykres dystrybuanty;

b) podać funkcję rozkładu zmiennej X.

c) Z jakim pp zmienna X przyjmuje wartości pomiędzy ¼ a 1?

Definicja (D4). Niech $f(x)$ - funkcja rozkładu zmiennej X, a statystyka Y oparta na zmiennej X ma postać $\varphi(X)$. Wówczas *wartością oczekiwaną (nadzieją matematyczną)* statystyki Y nazywamy liczbę $E(Y) := \int_{-\infty}^{+\infty} \varphi(x) f(x) dx \equiv \int_{0}^{1} \varphi(x) dF(x)$.

Wartość oczekiwana określa przeciętną (średnią) wartość zmiennej Y, jaką będziemy otrzymywać w kolejnych ekperymentach.

Bardzo ważny

Przykład (P2). Wartością oczekiwaną samej zmiennej *X* {tj. dla Y(X) = id(X) }, jest jej wartość *średnia,* $M(X) \equiv m_X = \int x\ f(x)\ dx$.

Dla dyskretnej zmiennej losowej (deltowej funkcji rozkładu), wzór ten przechodzi w zwykłą średnią z *N* liczb, będących wartościami X w każdym kolejnym punkcie od 1 do *N*.

Czytelnik odpowie, jakiej wartości należy oczekiwać przy rzucaniu kostką sześcienną? Tzn., gdyby jego zarobki były równe codziennemu wynikowi rzutu kostką, to ile złotych średnio dziennie by zarabiał? Odpowiedź proszę potwierdzić obliczeniem.

Kolejny, nieodzowny dla Czytelnika

Przykład (P3). Weźmy jeszcze statystykę $Y := (X - M)^2$.

E(Y) jest *dyspersją* D(X) vel *wariancją*, czyli kwadratem s^2 *odchylenia standardowego* s zmiennej X.

Widać, że oddalenie danej wartości od średniej o kolejne jednostki, zwiększa wariancję

coraz dramatyczniej. I jednakowo, czy to poniżej, czy powyżej średniej. Wariancja zawsze rośnie i jest dodatnia – w najlepszym wypadku, dokładnie równa 0 (w jakim? Co on oznacza dla wyników eksperymentu?).

Co więc, własnymi słowami Czytelnika, wedle Jego intuicji, określa wariancja wyników?

Dla zmiennej dyskretnej, w celu znalezienia D, należy obliczyć średnią M, odjąć ją od każdego wyniku, różnice takie podnieść do kwadratu, zsumować (otrzymujemy coś, co nazywa się Sumą Kwadratów, czyli SS) i na koniec podzielić przez *N*-1 (w wypadku próby) lub *N* (w wypadku całej populacji). Owo *N-1* jest niczym innym, jak *liczbą stopni swobody, df,* danej zmiennej X.

Średnia M mierzy to, gdzie wypada środek ciężkości rozkładu (wokół jakiego punktu na tarczy strzeleckiej koncentrują się nasze trafienia). Wariancja s^2 określa, jak nieostry, czyli rozproszony jest rozkład (czyli jak bardzo rozstrzelone są trafienia w tarczę).

Jeśli jakieś narzędzie badawcze charakteryzuje się skupianiem się wyników w innym miejscu, niż w tym, którego oczekiwaliśmy, to znaczy, że narzędzie nasze ma złą *trafność.*

Jeśli nasze narzędzie badawcze charakteryzuje się dużym rozstrzałem wyników, to znaczy, że ma niską *rzetelność.*

Dobrze skonstruowane eksperymenty muszą mieć zarówno dobrą trafność, jak i rzetelność.

Ćwiczenie (C3), czas: 20 minut. Mamy zbiór wartości dyskretnych zmiennej losowej (wyników naszego eksperymentu) w badaniu nr (1): 1, -1, 3, 2, 1, 4, -2, 0. W kolejnym badaniu, nr (2), wynikami były: 2, 1, 1, 0, -3, 1.

Z teorii wynika, że pomiary powinny oscylować wokół 1.

Która z serii badawczych była trafniejsza, a która bardziej rzetelna?