SdI30 W04: *PODSTAWY STATYSTYKI MATEMATYCZNEJ*

1. **Różne pojęcia statystyki**
   Przykład 1
2. **Badanie statystyczne**
3. **Populacja generalna i cecha statystyczna**
4. **Wnioskowanie statystyczne**
5. **Próba a próba reprezentatywna**
6. **Rozkład teoretyczny a rozkład empiryczny**
7. **Twierdzenie o rozkładzie średniej arytmetycznej**
   Przykład 2
8. **CTG – centralne twierdzenie graniczne**
   Przykład 3
9. **CTG dla sumy**

1. Różne pojęcia statystyki

A. ***Statystyka jako nauka*** dostarcza metod pozyskiwania, przetwarzania, zestawiania, analizy i prezentacji danych dotyczących wyników doświadczeń, obserwacji zjawisk losowych lub procesów masowych.

Wiele nauk zajmuje się badaniem „otaczającego nas świata" poprzez obserwacje lub konstrukcje doświadczeń dla potwierdzenia swoich teorii. Takie badania wymagają specjalistycznych metod i zwykle przebiegają według schematu:

- planowanie doświadczenia,
- zebranie i opracowanie danych,
- analiza danych, ich interpretacja i wnioski.

Statystyka tworzy i rozwija te metody w sposób formalny.

B. *Statystyka opisowa* (*descriptive statistics*) – zespół metod, nie używających probabilistyki, służących do wydobywania „informacji” zawartych w zbiorach danych zebranych w czasie ***badania statystycznego***, jako wyniku obserwacji, realizacji zjawiska lub doświadczenia losowego.

Celem stosowania metod statystyki opisowej jest podsumowanie zbioru danych i wyciągnięcie podstawowych wniosków dotyczących przedmiotu badań w określonej zbiorowości.

Przedmiotem zainteresowania statystyki opisowej są m.in.:
1. miary położenia: np. średnia, percentyle, wartość modalna.
2. miary dyspersji: np. wariancja, odchylenie standardowe,
3. miary asymetrii,
4. miary współzależności.

C. ***Statystyka matematyczna*** (SM) (*mathematical statistics*) – sformalizowany dział statystyki, używający probabilistyki i innych działów matematyki do badania poprawności przyjętych założeń, w określonym modelu probabilistycznym, na podstawie analizy danych otrzymanych w wyniku obserwacji zjawiska lub przeprowadzonego eksperymentu.

SM dostarcza teoretycznych podstaw do konstrukcji procedur statystycznych, w celu uzyskania wiarogodnej informacji o przedmiocie badania.

W SM wyniki doświadczenia zwane obserwacjami lub pomiarami, interpretujemy jako ciąg zm. l. $X_1, X_2, \ldots, X_n$ tworzących próbę losową $\boldsymbol{X}$. Zmienne te i ich rozkłady stanowią element modelu probabilistycznego badanego zjawiska.

D. Statystyka jako funkcja (*statistic*) – każda zm. l. $U = f(\boldsymbol{X}_n)$ będąca funkcją próby losowej.

Statystyki służą do poznania mechanizmu generującego obserwacje.

Dzięki probabilistyce znamy m. in. twierdzenia, których tezy dotyczą rozkładu najczęściej stosowanych statystyk.

Podstawowe statystyki:

- minimalna statystyka: $X_{(1)} = \min(\boldsymbol{X})$,
- maksymalna statystyka: $X_{(n)} = \max(\boldsymbol{X})$,
- rozstęp: $R = X_{(n)} - X_{(1)}$,
- suma: $T_n = \sum_{i=1}^{n} X_i$,
- średnia arytmetyczna: $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

- jeżeli modelem cechy jest zm. l. $X \sim B(p)$, to średnią arytmetyczną nazywamy frakcją jednostek wyróżnionych w próbie i oznaczamy $\bar{P}_n$.
- wariancja z próby: $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})^2$,
- odchylenie standardowe z próby: $S_n = +\sqrt{S_n^2}$,
- kowariancja empiryczna:

  $\mathrm{Cov}(\boldsymbol{X}_n, \boldsymbol{Y}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}}_n)(Y_i - \bar{\boldsymbol{Y}}_n)$ dla $n \geq 2$,
- współczynnik korelacji Pearsona:

$$\mathrm{Corr}(\boldsymbol{X}_n, \boldsymbol{Y}_n) = \frac{\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}}_n)(Y_i - \bar{\boldsymbol{Y}}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}}_n)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{\boldsymbol{Y}}_n)^2}},$$

Wartości statystyk oznaczamy małymi literami.

Przykład 1. Ze zbioru $\{1, 2, 3, 4\}$ losowana jest dwuelementowa próba bez zwracania. Wyznaczyć współczynnik korelacji pomiędzy statystykami $X_{(1)}$ i $X_{(2)}$.

$\Omega = (X_1, X_2)$	$\Pr\{(x_1, x_2)\}$	$X_{(1)}$	$X_{(2)}$
$(1, 2)$	$1/12$	1	2
$(1, 3)$	$1/12$	1	3
$(1, 4)$	$1/12$	1	4
$(2, 1)$	$1/12$	1	2
$(2, 3)$	$1/12$	2	3
$(2, 4)$	$1/12$	2	4
$(3, 1)$	$1/12$	1	3
$(3, 2)$	$1/12$	2	3
$(3, 4)$	$1/12$	3	4
$(4, 1)$	$1/12$	1	4
$(4, 2)$	$1/12$	2	4
$(4, 3)$	$1/12$	3	4

Rozkład łączny

$x_{(1)} \backslash x_{(2)}$	2	3	4	$f_{X_{(1)}}(x_{(1)})$
1	$2/12$	$2/12$	$2/12$	$6/12$
2	0	$2/12$	$2/12$	$4/12$
3	0	0	$2/12$	$2/12$
$f_{X_{(2)}}(x_{(2)})$	$2/12$	$4/12$	$6/12$	$12/12$

Rozkład brzegowy $X_{(1)}$

$X_{(1)}$	1	2	3
$f_{X_{(1)}}(x_{(1)})$	$6/12$	$4/12$	$2/12$

Rozkład brzegowy $X_{(2)}$

$X_{(2)}$	2	3	4
$f_{X_{(2)}}(x_{(2)})$	$2/12$	$4/12$	$6/12$

$$\mathbb{E}X_{(1)} = \frac{5}{3}, \qquad \mathbb{E}X_{(2)} = \frac{10}{3}$$

$$\mathbb{E}X_{(1)}^2 = \frac{10}{3}, \qquad \mathbb{E}X_{(2)}^2 = \frac{35}{3}$$

$$\mathbb{D}^2X_{(1)} = \frac{5}{9}, \qquad \mathbb{D}^2X_{(2)} = \frac{5}{9}$$

$$\mathbb{E}\left(X_{(1)}X_{(2)}\right) = \frac{35}{6}$$

$$\operatorname{cov}\left(X_{(1)}, X_{(2)}\right) = \frac{5}{18}$$

$$\operatorname{corr}\left(X_{(1)}, X_{(2)}\right) = \frac{\operatorname{cov}\left(X_{(1)}, X_{(2)}\right)}{\sqrt{\mathbb{D}^2X_{(1)} \cdot \mathbb{D}^2X_{(2)}}} = \frac{1}{2}$$

2. Badanie statystyczne

Badanie statystyczne (BS) to szereg czynności związanych z pozyskiwaniem i przetwarzaniem danych zmierzających do jak najlepszego poznania rozkładu badanych cech statystycznych $X, Y, \ldots, Z$ w ***populacji***.

Badanie może być:

- pełne – obejmuje całą populację,
- częściowe – dotyczy wyodrębnionych elementów populacji, tworzących ***próbę***.

Czynniki, które przemawiają na korzyść badań częściowych:

- populacja może być nieskończona,
- badanie może być niszczące,
- wysokie koszty.

3. Populacja generalna i cecha statystyczna

Populacja generalna (zbiorowość statystyczna) to zbiór elementów zwanych ***jednostkami statystycznymi***, podlegających badaniu. Jednostki populacji są do siebie podobne pod względem badanych cech, ale nie są identyczne. Zróżnicowanie wartości cech często jest celem identyfikacji ich rozkładu.

Cechy statystyczne to te właściwości populacji, które są przedmiotem badań. Modelami badanych cech statystycznych są zmienne losowe.

Cecha statystyczna może być:

- ***mierzalna*** (*ilościowa*), np. długość detalu, czas zdatności obiektu, pojemność kondensatora, szybkość procesora;
- ***niemierzalna*** (*jakościowa*), np. kolor oczu, nazwa firmy, upodobania klienta.

4. Wnioskowanie statystyczne

Wnioskowanie statystyczne to metody uogólniania wyników badań próby losowej na całą populację oraz szacowania błędów wynikających z takiego uogólnienia.

Wyróżniamy dwie grupy metod uogólniania wyników, tworzące zarazem dwa działy wnioskowania statystycznego:

- ***Estymacja*** – szacowanie wartości nieznanych parametrów lub postaci rozkładów badanych cech.
- ***Weryfikacja hipotez statystycznych*** – sprawdzanie poprawności przypuszczeń dotyczących parametrów lub postaci rozkładu badanych cech, związków pomiędzy badanymi cechami w jednej lub kilku populacjach, losowości próby i innych specyficznych właściwości statystycznych.

5. Próba a próba reprezentatywna

Próbą losową (*random sample*) z populacji badanej ze względu na jedną cechę X, lub kilka cech, np. dwie X i Y nazywamy:

- w przypadku jednej cechy ciąg zm. l.

$$X_1, X_2, \ldots, X_n \text{ oznaczany } \boldsymbol{X} \text{ lub } \boldsymbol{X}_n,$$

- w przypadku dwóch cech ciąg par zm. l.

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \text{ oznaczany } (\boldsymbol{X}, \boldsymbol{Y})$$

każda z określonym rozkładem prawdopodobieństwa.

Jeżeli badamy dwie populacje ze względu na wspólną cechę X, to próbą losową są dwa ciągi: $X_{1,1}, \ldots X_{1,n}$ i $X_{2,1}, \ldots, X_{2,m}$.

Jeżeli zm. l.-owe w próbie są niezależne i o identycznym rozkładzie co badana cecha lub cechy, to próbę nazywamy ***prostą próbą losową*** i ozn. SRS (*simple random sample*).

Próbą reprezentatywną nazywamy taką próbę, która zachowuje strukturę populacji ze względu na badane cechy.

SRS gwarantuje reprezentatywność.

Próbę niereprezentatywną nazywamy ***próbą obciążoną***.

Planowaniem doświadczenia i sposobem wyboru próby zajmuje się dział statystyki zwany ***metody reprezentacyjne***.

Czesław Bracha. *Teoretyczne podstawy metody reprezentacyjnej*. Wyd. Naukowe PWN.

Liczbę n jednostek wybranych do próby nazywamy ***licznością próby***. Liczność próby zależy m.in. od przyjętego błędu, zwanego ***poziomem ufności***.

Jeżeli $n \leq 30$ to próbę nazywamy ***małą próbą***.

W przeciwnym przypadku próbę nazywamy ***dużą próbą***.

6. Rozkład teoretyczny a rozkład empiryczny

Probabilistycznym modelem badanej cechy jest zm. l. X. Rozkład cechy X w populacji nazywamy ***rozkładem teoretycznym***. Rozkład ten zwykle nie jest znany i w badaniach statystycznych zwykle przyjmujemy, że jest to pewien rozkład spośród określonej rodziny rozkładów, zależnej od nieznanych parametrów, np.

$$X \sim \mathcal{N}(\mu = ?, \sigma = ?),\ Y \sim \mathrm{B}(p = ?),\ T \sim \mathrm{bin}(n, p = ?).$$

Rozkład cechy lub kilku cech w próbie nazywamy ***rozkładem empirycznym***. Rozkład ten poznajemy na podstawie BS opisującego wartości przyjmowane przez cechę lub cechy, zwykle przy pomocy dystrybuanty empirycznej, częstości ich występowania lub innych odpowiednich statystyk z próby.

Niech $(X_1, X_2, \ldots, X_n)$ będzie jedno-cechową próbą prostą.

Dystrybuantą empiryczną (*empirical distribution function*) nazywamy funkcję $\widehat{F_n}$ określoną wzorem:

$$\forall_{x\in\mathbb{R}} \widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\},$$

gdzie $\mathbb{I}\{A\}$ jest indykatorem zdarzenia A. Dla ustalonego x,

$$\mathbb{I}\{X_i \leq x\} \sim \mathrm{B}\big(F(x)\big),$$

$$n\widehat{F_n}(x) \sim \mathrm{bin}(nF(x), nF(x)(1-F(x))).$$

Dystrybuanta empiryczna $\widehat{F_n}(x)$ jest nieobciążonym estymatorem dystrybuanty $F(x)$.

UWAGA. Niektóre pojęcia odnoszą się zarówno do populacji, jak i do próby. Rozróżniamy więc wariancję i odchylenie stand. cechy w populacji od ich odpowiedników w próbie.

7. Twierdzenie o rozkładzie średniej arytmetycznej

Jeżeli cechę w populacji generalnej opisuje zm. l. X o rozkładzie normalnym $\mathcal{N}(\mu, \sigma)$, to średnia arytmetyczna $\bar{\boldsymbol{X}}_n$ z SRS ma rozkład normalny $\mathcal{N}(\mu, \sigma/\sqrt{n})$, tj.

$$\underbrace{X \sim \mathcal{N}(\mu, \sigma)}_{założenie} \Rightarrow \underbrace{\bar{\boldsymbol{X}}_n \sim \mathcal{N}(\mu, \sigma/\sqrt{n})}_{teza}$$

Dowód tego twierdzenia wynika z lematu o rozkładzie sumy.

Lemat o rozkładzie sumy. Jeśli X_i, $i = 1, \ldots, n$ są niezależnymi zm. l. o rozkładach normalnych $\mathcal{N}(\mu_i, \sigma_i)$, to

$$(X_1 + \cdots + X_n) \sim \mathcal{N}\left(\mu_1 + \cdots + \mu_n,\ \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}\right)$$

Dowód indukcyjny. Najpierw wykażemy lemat dla $n = 2$. Niech $X_1 = U + \mu_1, \quad X_2 = W + \mu_2,$ gdzie $U \sim \mathcal{N}(0;\ \sigma_1)$, $W \sim \mathcal{N}(0;\ \sigma_2)$ oraz U i W są niezależne. Wystarczy wykazać, że $(U + W) \sim \mathcal{N}(0;\ \sqrt{\sigma_1^2 + \sigma_2^2})$. Stosujemy splot dla gęstości

$$f_{U+W}(x) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma_1^2}(x-y)^2 - \frac{1}{2\sigma_2^2}y^2\right) dy$$

Wprowadzamy $a^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$, $\sigma^2 = \sigma_1^2 + \sigma_2^2$, wówczas

$$f_{U+W}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \frac{1}{a\sqrt{2\pi}} \exp\left(-\frac{1}{2a^2}\left(y - \frac{a^2}{\sigma_1^2}x\right)^2\right) dy = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Po zastosowaniu metody indukcji matematycznej otrzymujemy tezę lematu.

Dowód z zastosowaniem funkcji charakterystycznych

Założenie. $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$,

Funkcja charakterystyczna rozkładu normalnego

$$\varphi(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-(x-\mu)^2/2\sigma^2}\, dx$$

Po przekształceniach otrzymujemy

$$\varphi(t) = e^{i\mu t - \sigma^2 t^2/2}$$

A w szczególności, możemy otrzymać funkcję charakterystyczną standardowego rozkładu normalnego $\mathcal{N}(0, 1)$,

$$\varphi(t) = e^{-t^2/2}$$

Stąd funkcja charakterystyczna sumy $X_1 + X_2$

$$\varphi(t) = e^{i\mu_1 t - \frac{\sigma_1^2 t^2}{2}} \cdot e^{i\mu_2 t - \frac{\sigma_2^2 t^2}{2}} = e^{i(\mu_1+\mu_2)t - \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}$$

Zatem

$$(X_1 + X_2) \sim \mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Stąd wynika dowód twierdzenia.

Jeżeli $X_i, X_2, \ldots, X_n$ jest SRS z rozkładu $\mathcal{N}(\mu, \sigma)$, to

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, \sqrt{\sum_{i=1}^{n} \sigma_1^2}\right)$$

Wniosek z twierdzenia.

Ponieważ $\overline{\boldsymbol{X}}_n \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, więc po standaryzacji średniej:

$$\frac{\overline{\boldsymbol{X}}_n - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$$

Uwaga. W statystyce stosowanej twierdzenia probabilistyki postaci $\alpha \Rightarrow \beta$ są stosowane w drugą stronę, tzn. z pewnej wiedzy zawartej w tezie β wnioskujemy o prawdziwości założenia α. Wnioskowanie to nazywamy ***wnioskowaniem redukcyjnym***, w odróżnieniu od dedukcyjnego dowodzenia prawdy stosowanego w naukach formalnych.

Wnioskowanie redukcyjne nie jest niezawodne, niemniej jest najczęściej stosowane w naukach empirycznych.

Przykład 2. Długość linii jaką można narysować pewnego typu pisakiem ma rozkład $\mathcal{N}(800; 100)[m]$.

a) Ile trzeba mieć takich pisaków, aby z prawd. co najmniej 0,99, można było narysować linię o długości ponad $3000[m]$?

b) Co wynika z faktu, że średnia długość linii narysowanej 4 pisakami jest krótsza niż $650[m]$?

https://pl.wikisource.org/wiki/Tablica_rozkładu_normalnego

http://davidmlane.com/hyperstat/z_table.html

8. CTG – centralne twierdzenie graniczne (CLT)

Jeżeli $\boldsymbol{X}_n$, $n = 1, 2, ...$, jest ciągiem prostych prób losowych (SRS) z populacji, w której badana cecha X ma skończoną wartość oczekiwaną $\mathbb{E}X$ oraz dodatnie i skończone odchylenie standardowe $\mathbb{D}X$, to wraz ze wzrostem liczebność prób, rozkład ciągu średnich arytmetycznych $\bar{\boldsymbol{X}}_n$, dąży do rozkładu normalnego z parametrami $\mathbb{E}X$ i $\mathbb{D}X/\sqrt{n}$, tj.

$$\underbrace{X \sim ?\,(\text{znane: } \mathbb{E}X, \mathbb{D}X)}_{założenie} \Rightarrow \underbrace{\bar{\boldsymbol{X}}_n \underset{n\to\infty}{\sim} \mathcal{N}(\mathbb{E}X, \mathbb{D}X/\sqrt{n})}_{teza}$$

Siła CTG polega na tym, że rozkład populacji może być inny niż normalny, a nawet może być nieznany, stąd znak „?”.

Twierdzenie o standaryzowanym rozkładzie średniej arytmetycznej nazywa się ***tw. Lindeberga-Levy'ego***.

Przykład 3. Dane techniczne informują, że pewne silniki osiągają średni max moment obrotowy 220[Nm], a odchylenie standardowe wynosi 15[Nm]. Producent łodzi motorowych zanim dokona zakupu tych silników zamierza zbadać próbną partię 36 silników. Jakie jest prawd. zdarzenia, że średni max moment z próby przyjmie wartość mniejszą niż 215[Nm]? Jeśli średni moment z próby będzie mniejszy od 215[Nm], to jaki stąd wyciągniemy wniosek?

Badana cecha X jest maksymalnym momentem obrotowym silnika.

Dane: $\mathbb{E}X = 220[\mathrm{Nm}]$, $\mathbb{D}X = 15[\mathrm{Nm}]$, $n = 36$, więc próba jest duża, nie znamy rozkładu zm. l. X.

Szukane: $P(\overline{\mathbf{X}}_{36} < 215) = ?$

Na mocy CTG średnia arytmetyczna z próby $\overline{\mathbf{X}}_{36}$ ma w przybliżeniu rozkład normalny z parametrami:

$$\mathbb{E}\overline{\boldsymbol{X}}_{36} = \mathbb{E}X = 220 \text{ i } \mathbb{D}\overline{\boldsymbol{X}}_{36} = \frac{\mathbb{D}X}{\sqrt{n}} = 2{,}5.$$

Aby skorzystać z tablicy wartości dystrybuanty Φ dokonujemy standaryzacji średniej arytmetycznej

$$\mathrm{P}(\overline{\mathbf{X}}_{36} < 215) \overset{STD}{\approx} \mathrm{P}\left(Z < \frac{215-\mathbb{E}X}{\mathbb{D}X/\sqrt{n}}\right) = \mathrm{P}\left(Z < \frac{215-220}{2{,}5}\right) =$$

$$\Phi(-2) \overset{TABL}{=} 0{,}0228$$

Wniosek. Prawd. że test, który chce przeprowadzić nabywca, wykaże średni max moment obrotowy silnika mniejszy niż 215[Nm] jest bardzo małe. Wynika stąd, że jeśli przeprowadzony test da wynik mniejszy od 215[Nm], to będą podstawy do podważenia danych technicznych silnika, tj. a priori danej informacji o parametrach osiąganej mocy silników.

9. CTG dla sumy

Jeżeli $\boldsymbol{X}_n$ jest SRS z populacji X o skończonej wartości oczekiwanej $\mathbb{E}X$ i odchyleniu stand. $\mathbb{D}X$, to

$$\sum_{i=1}^{n} X_i \underset{n\to\infty}{\sim} \mathcal{N}(n\mathbb{E}X, \sqrt{n}\,\mathbb{D}X)$$

Dowód. Spełnione są założenia CTG, więc $\overline{\boldsymbol{X}}_n \sim \mathcal{N}\left(\mathbb{E}X, \frac{\mathbb{D}X}{\sqrt{n}}\right)$.

Ponieważ $X_1 + \cdots + X_n = n\overline{\boldsymbol{X}}_n$, więc suma $n\overline{\boldsymbol{X}}_n$ ma asymptotycznie rozkład normalny z wartością oczekiwaną

$$\mathbb{E}(n\overline{\boldsymbol{X}}_n) = n\mathbb{E}\overline{\boldsymbol{X}}_n = n\mathbb{E}X$$

i wariancją

$$\mathbb{D}^2(n\overline{\boldsymbol{X}}_n) = n^2\mathbb{D}^2\overline{\boldsymbol{X}}_n = n^2\,\frac{\mathbb{D}^2X}{n}, \text{ stąd } \mathbb{D}(n\overline{\boldsymbol{X}}_n) = \sqrt{n}\,\mathbb{D}X.$$

Przykład 4. Dzienne obroty pewnego sklepu z częściami samochodowymi są zmienną losową o wartości oczekiwanej 5500 zł i odchyleniu standardowym 1050 zł. Wyznaczyć prawdopodobieństwa następujących zdarzeń:

a) obroty w okresie 100 dni przekroczą łącznie 570 000[zł],

b) średnia z dziennych obrotów, realizowanych w okresie 100 dni będzie należała do przedziału $(5200;\ 5600)$[zł].

Odp.: a) 0,0287; b) 0,8268.

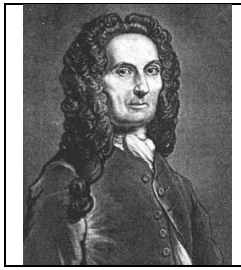
10. Twierdzenie de Moivre'a-Laplace'a

Tw. de Moivre'a[1]-Laplace'a ma zastosowanie w statystyce. Jest szczególnym przypadkiem CTG i dotyczy rozkładu frakcji $\bar{P}_n$ lub sumy T_n ciągu $X_1, X_2, \ldots, X_n$ zm. l. stanowiących SRS o rozkładzie $B(p)$.

Gdy liczność n ciągu zm. losowych rośnie, to częstość

$$\bar{P}_n = \frac{1}{n} T_n, \text{ gdzie } T_n = \sum_{i=1}^{n} X_i,$$

zwana w statystyce ***frakcją z próby*** ma rozkład zbieżny do rozkładu normalnego z parametrami:



[1] Abraham de Moivre (1667 – 1754) was a French-born mathematician who pioneered the development of analytic geometry and the theory of probability.

$$\mathbb{E}\bar{P}_n = p \text{ i } \mathbb{D}^2\bar{P}_n = \frac{p(1-p)}{n},$$

$$\text{czyli } \bar{P}_n \underset{n\to\infty}{\sim} \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Ponadto suma T_n jest również asymptotycznie normalna, tj.

$$\underbrace{T_n \sim \text{bin}(n,p)}_{założenie} \Longrightarrow \underbrace{T_n \underset{n\to\infty}{\sim} \mathcal{N}\left(np, \sqrt{np(1-p)}\right)}_{teza}$$

Jeśli $p = 0{,}5$, to suma T_n ma symetryczny rozkład dwumianowy i zbieżność do rozkładu normalnego jest bardzo szybka. Jeśli parametr p jest bliski 0 (lub 1), to rozkład dwumianowy jest silnie asymetryczny, ale ze wzrostem n asymetria zanika.

W praktyce przyjmujemy, że przybliżenie rozkładu dwumianowego rozkładem normalnym jest dobre, gdy liczebność n jest na tyle duża, że spełniony jest warunek:

$$0 < \mathbb{E}X \pm 3\mathbb{D}X < n,$$

czyli wartości $np \pm 3\sqrt{np(1-p)}$ należą do przedziału $(0, n)$.

Uwaga. Korzystając dla rozkładu dwumianowego z przybliżenia rozkładem normalnym należy uwzględnić ***poprawkę na ciągłość*** tj., jeżeli $X \sim \mathrm{bin}(n, p)$, to

$$\mathrm{P}(X \leq a) \approx \Phi\left(\frac{(a+0{,}5)-np}{\sqrt{np(1-p)}}\right), \quad \mathrm{P}(X \geq a) \approx 1 - \Phi\left(\frac{(a-0{,}5)-np}{\sqrt{np(1-p)}}\right),$$

$$\mathrm{P}(a \leq X \leq b) \approx \Phi\left(\frac{(b+0{,}5)-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{(a-0{,}5)-np}{\sqrt{np(1-p)}}\right).$$

Przykład 5. Niech $X \sim \text{bin}(10;\ 0{,}5)$. Obliczyć dokładną i aproksymowaną wartość prawd. zdarzenia $X \leq 4$.

 Obliczamy $\text{P}(X \leq 4)$ dla rozkładu $\text{bin}(10;\ 0{,}5)$

$$\text{P}(X \leq 4) = F_{\text{bin}}(4|10; 0{,}5) = \sum_{x=0}^{4} f_{\text{bin}}\,(x|10; 0{,}5) = 0{,}377$$

Wartości $\mathbb{E}X - 3\mathbb{D}X \approx 0{,}25$ i $\mathbb{E}X + 3\mathbb{D}X = 9{,}75$ leżą między liczbami 0 i $n = 10$, więc rozkład $\mathcal{N}\left(5; \sqrt{2{,}5}\right)$ jest dobrym przybliżeniem dla rozkładu $\text{bin}(10;\ 0{,}5)$.

Zauważmy jednak, że $F_{\mathcal{N}}(4|5; 1{,}581) = 0{,}263394$ różni się znacząco od właściwego wyniku. Przyczyną dużej różnicy jest nieuwzględnienie poprawki na ciągłość. Należy więc dodać 0,5 do 4, zanim przystąpimy do obliczenia przybliżonej wartości prawd.

Porównajmy wyniki

$X \sim \mathrm{bin}(10;\ 0{,}5)$	$Z \sim \mathcal{N}(5;\ 1{,}581)$
$F_X(4) = 0{,}376953$	$F_Z(4{,}5) = 0{,}375826$
$F_X(5) = 0{,}623047$	$F_Z(5{,}5) = 0{,}624174$
$F_X(6) = 0{,}828125$	$F_Z(46{,}5) = 0{,}828784$

Jak widać, uwzględniając poprawkę otrzymujemy całkiem dobre przybliżenia.

Przykład 6. Jak liczną próbę należy pobrać z populacji, w której obserwowana cecha X ma rozkład $B(0{,}4)$, aby z prawd. co najmniej 0,9 można było stwierdzić, że częstość sukcesów będzie odchylać się od prawd. sukcesu w jednym doświadczeniu nie więcej niż o 0,1?

 Niech $X_1, \dots, X_n$ będzie SRS oraz $X_i \sim B(p = 0{,}4)$ dla $i = 1, \dots, n$, gdzie n należy wyznaczyć.

Częstość sukcesów wyraża się wzorem:

$$\bar{P}_n = \frac{1}{n} T_n,$$

gdzie zm. l. $T_n = X_1 + \cdots + X_n$ zlicza sukcesy.

Liczbę n otrzymamy z nierówności $\mathrm{P}(|\bar{P}_n - 0{,}4| \leq 0{,}1) \geq 0{,}9$ i tw. M-L'a (bez uwzględniania poprawki na ciągłość)

$$\mathrm{P}(-0{,}1n \leq T_n - 0{,}4n \leq 0{,}1n) \overset{STD}{=} \mathrm{P}\left(\frac{-0{,}1n}{\sqrt{0{,}24n}} \leq Z \leq \frac{0{,}1n}{\sqrt{0{,}24n}}\right)$$

$$= 2\Phi\left(\frac{0{,}1n}{\sqrt{0{,}24n}}\right) - 1 \geq 0{,}9 \Leftrightarrow \frac{0{,}1n}{\sqrt{0{,}24n}} \geq \Phi^{-1}(0{,}95) \overset{TABL}{=} 1{,}64,$$

$$\text{stąd } n \geq 65.$$

11. Zestaw zadań W04

1. Zużycie wody (w hektolitrach) w pewnym osiedlu w ciągu dnia ma rozkład $\mathcal{N}(m = ?, \sigma = 11)$. Obliczyć prawd. zdarzenia, że empiryczna wariancja zużycia wody w losowo wybranych 90 dniach

a) nie będzie większa niż 100[hl],
b) będzie większa niż 200[hl].

2. Wiadomo, że błąd pomiaru pewnego przyrządu ma rozkład normalny $\mathcal{N}(0, \sigma)$ i z prawd. 0,95 nie wychodzi poza przedział $(-1, 1)$. Dokonanych zostanie i) 10, ii) 100 niezależnych pomiarów tym przyrządem. Oblicz prawd. zdarzenia, że wariancja pomiarów

a) przyjmie wartość między 0,2 a 0,3,
b) będzie większa od 0,28.

Odp. i) a) 0,1665, ii) b) $\approx$ 0,27.

3. Losujemy 100 liczb według rozkładu jednostajnego na przedziale $(0, 1)$.

a) Ustalić rozkład sumy tych liczb.

b) Obliczyć prawd. zdarzenia, że suma wylosowanych liczb nie będzie należała do przedziału $(45, 55)$.

c) Wyznaczyć dystrybuantę największej z wylosowanych liczb i oblicz prawd., że liczba ta będzie mniejsza od 0,95.

d) Jaki wniosek należy wyciągnąć, jeśli że suma wylosowanych liczb będzie mniejsza niż 40?

4. Niech $X_1, \dots, X_n$ będzie próbą prostą z populacji, w której cecha X ma rozkład o gęstości

$$f(x) = \frac{x}{8}\mathbf{1}_{(0;\ 4)}(x)$$

a) Wyznaczyć dystrybuantę i gęstość statystyk

$$Y = \max\{X_1, \dots, X_n\}, Z = \min\{X_1, \dots, X_n\},$$

b) Obliczyć prawd. zdarzeń $Y < 3, Z > 1$.

c) Obliczyć wartości oczekiwane i wariancje Y i Z.

5. Ze zbioru $\{1, 2, 3, 4\}$ wylosowano dwie liczby

i) ze zwracaniem,

ii) bez zwracania,

Wyznaczyć rozkład oraz wartość oczekiwaną i wariancję rozstępu.