

# Statystyka w analizie i planowaniu eksperymentu

Paweł Błażej

6 marca 2013

Zdarzeniom losowym określonym na pewnej przestrzeni zdarzeń elementarnych  $\Omega$  można zazwyczaj na wiele różnych sposobów przypisać jakieś prawdopodobieństwa. Trójka  $(\Omega, \mathcal{F}, \mathcal{P})$ , gdzie  $\mathcal{P}$  jest rodziną funkcji prawdopodobieństwa nazywamy **przestrzenią statystyczną**.

Przestrzeń statystyczna służy do opisu możliwych mechanizmów rządzących eksperymentem losowym.

Wybór przestrzeni statystycznej jest zazwyczaj kompromisem między prostotą modelu a jego adekwatnością.

Oznaczmy przez  $p$  prawdopodobieństwo otrzymania orła przy jednokrotnym rzucie monetą. Wówczas

$$\Omega = \{O, R\}, \mathcal{F} = \{\emptyset, \{O\}, \{R\}, \{O, R\}\}$$

oraz

$$\mathcal{P} = \{p : 0 \leq p \leq 1\}$$

## Po co ta statystyka?

**Jednym z podstawowych zadań statystyki jest podanie metod, które umożliwiają identyfikację rozkładu prawdopodobieństwa, rządzącego eksperymentem losowym, na podstawie obserwacji jego wyniku  $X$ .**

W dalszej części wykładu zostanie pokazane, że wielokrotne powtarzanie eksperymentu może znacznie ułatwić identyfikację opisującego go rozkładu prawdopodobieństwa.

Zebrane wyniki eksperymentu np.  $X_1, X_2, \dots, X_n$   
(gdzie  $X_1, X_2, \dots$  (zmiennie losowe)) nazywamy **próbą losową**

Jeżeli rozkłady zmiennych losowych  $X_1, X_2, \dots, X_n$  są niezależne, to mamy do czynienia z **prostą próbą losową**.

# Przykłady prób losowych

- 1 wyniki rzutów monetą;
- 2 wyniki rzutów kostką;
- 3 czas obsługi klientów przy kasie;
- 4 długość piór pawia.

Z każdą próbą losową można związać funkcję  $T$  o wartościach rzeczywistych, każdą taką funkcję nazywamy **statystyką**



- 1  $\bar{X}$
- 2  $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
- 3  $S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- 4  $mo, me, d, X_{min}, X_{max}$ .

## Średnia z próby

Niech  $X_1, X_2, \dots, X_n$  będzie próbą losową. Średnią z próby nazywamy statystykę

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

## Uwaga

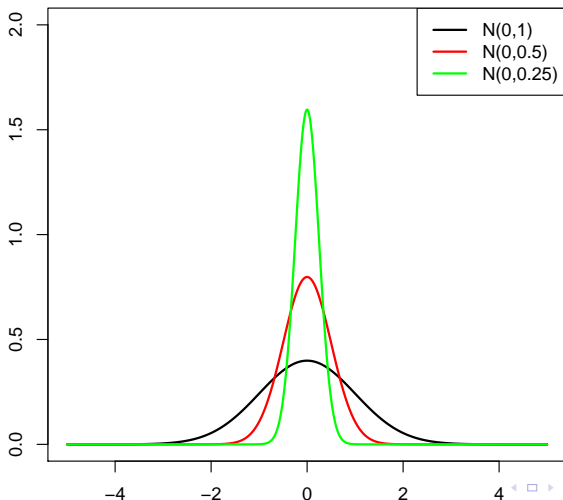
Średnia jako funkcja próby losowej jest zmienną losową

- 1 Jeżeli  $EX_1 = \dots = EX_n = \mu$ , to  $E\bar{X}_n = \mu$

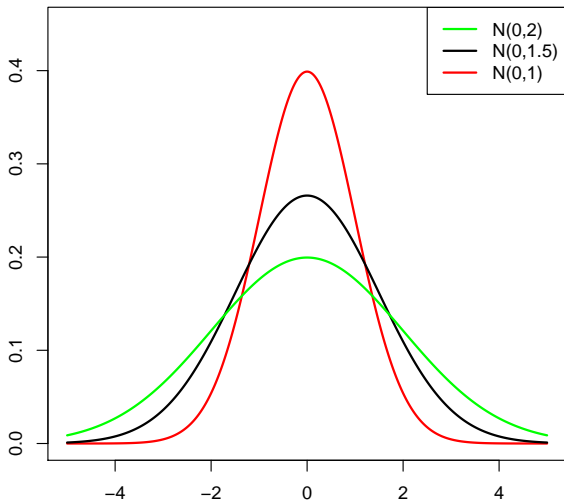
- 1 Jeżeli  $EX_1 = \dots = EX_n = \mu$ , to  $E\bar{X}_n = \mu$
- 2 Jeżeli  $X_1, \dots, X_n$  jest próbą prostą oraz  $EX_i = \mu$  i  $VarX_i = \sigma^2 < \infty$  dla  $i = 1, 2, \dots, n$ , to  $Var\bar{X}_n = Var(\frac{1}{n} \sum_{i=1}^n X_i) = \frac{1}{n}\sigma^2$

- 1 Jeżeli  $EX_1 = \dots = EX_n = \mu$ , to  $E\bar{X}_n = \mu$
- 2 Jeżeli  $X_1, \dots, X_n$  jest próbą prostą oraz  $EX_i = \mu$  i  $VarX_i = \sigma^2 < \infty$  dla  $i = 1, 2, \dots, n$ , to  $Var\bar{X}_n = Var(\frac{1}{n} \sum_{i=1}^n X_i) = \frac{1}{n}\sigma^2$
- 3 Jeżeli  $X_1, X_2, \dots, X_n$  są zmiennymi losowymi z rozkładu normalnego  $N(\mu, \sigma)$ , to  $\bar{X}$  jest zmienną losową o rozkładzie  $N(\mu, \sigma/\sqrt{n})$

# Rozkład normalny własności średniej z próby - interpretacja



# Rozkład normalny interpretacja parametru skali



Niech  $X$  będzie zmienną losową taką, że  $EX = \mu$  i o skończonej wariancji i niech  $X_1, \dots, X_n$  będzie prostą próbą losową z rozkładu zmiennej  $X$ . Wówczas

$$\bar{X}_n \rightarrow E[X] \text{ gdy } n \rightarrow \infty.$$



Nich  $X_1, X_2, \dots, X_n$ , będą to wyniki kolejnych rzutów monetą. Z prawa wielkich liczb wiemy, że  $\bar{X} \rightarrow E[X] = p$ .

Stąd średnia arytmetyczna z próby jest oszacowaniem (estymatorem) parametru  $p$ .

Wygeneruj  $n$ -krotny rzut monetą dla  $n = 50, 500, 5000$ . Oblicz średnią dla każdego z tych przypadków a następnie porównaj wyniki.

```
[1] "n=50"
```

```
[1] 0.5
```

```
[1] "n=500"
```

```
[1] 0.518
```

```
[1] "n=5000"
```

```
[1] 0.491
```

## Uwaga

Prawo wielkich liczb nie daje żadnej odpowiedzi na pytanie jak liczna powinna być próba aby przybliżenie nieznannej wartości oczekiwanej było „dobre”.

W praktyce zdarza się, że zamiast szacowania wartości nieznanego parametru (przy pomocy estymatorów) decydujemy się na podanie przedziału, do którego z pewnym prawdopodobieństwem należy szacowany przez nas parametr.

## Problem

Oszacowanie parametru  $\mu$  na podstawie próby  $X_1, X_2, \dots, X_n$  z rozkładu  $\mathcal{N}(\mu, \sigma)$ .

## Problem

Oszacowanie parametru  $\mu$  na podstawie próby  $X_1, X_2, \dots, X_n$  z rozkładu  $\mathcal{N}(\mu, \sigma)$ .

## Uwaga

Rozpatrujemy dwa przypadki

- 1 parametr  $\sigma$  **jest znany** (rzadko spotykany w praktyce);
- 2 parametr  $\sigma$  **nie jest znany** (przypadek często spotykany w praktyce).

# Przypadek nr. 1

Niech  $X_1, X_2, \dots, X_n$  będzie to próba losowa z rozkładu  $N(\mu, \sigma)$  przy czym parametr  $\mu$  jest **nieznany** natomiast parametr  $\sigma$  jest **znany**.

Po serii łatwych przekształceń :) otrzymujemy:

$$P(\bar{X} - a_\alpha \sigma / \sqrt{n} \leq \mu \leq \bar{X} + a_\alpha \sigma / \sqrt{n}) = 1 - \alpha.$$

Jest to przedział ufności dla parametru  $\mu$  utworzony na podstawie próby losowej  $X_1, X_2, \dots, X_n$  na poziomie istotności  $1 - \alpha$



# Przypadek nr. 1

Niech  $X_1, X_2, \dots, X_n$  będzie to próba losowa z rozkładu  $N(\mu, \sigma)$  przy czym parametr  $\mu$  jest **nieznany** natomiast parametr  $\sigma$  jest **znany**.

Po serii łatwych przekształceń :) otrzymujemy:

$$P(\bar{X} - a_\alpha \sigma / \sqrt{n} \leq \mu \leq \bar{X} + a_\alpha \sigma / \sqrt{n}) = 1 - \alpha.$$

Jest to przedział ufności dla parametru  $\mu$  utworzony na podstawie próby losowej  $X_1, X_2, \dots, X_n$  na poziomie istotności  $1 - \alpha$

## Uwaga

$a_\alpha$  jest to wartość kwantyla (rzędu  $1 - \alpha/2$ ) z rozkładu normalnego  $\mathcal{N}(0, 1)$ .

## Przypadek nr. 2

Niech  $X_1, X_2, \dots, X_n$  będzie to próba losowa z rozkładu  $\mathcal{N}(\mu, \sigma)$  przy czym parametr  $\mu$  oraz parametr  $\sigma$  jest **nieznany**. Przypadek ten jest nieco trudniejszy i wymaga „szczypty” wiedzy matematycznej. Podzielimy ten przypadek na dwa tzn. Przypadek nr .2a oraz Przypadek nr. 2b.

Niech  $X_1, X_2, \dots, X_n$  gdzie  $n \leq 30$  będzie to próba losowa z rozkładu  $N(\mu, \sigma)$  przy czym parametr  $\mu$  oraz parametr  $\sigma$  jest **nieznany**.

## Przypadek 2a „szczypta” matematyki

Niech  $X_1, X_2, \dots, X_n$  będzie próbą prostą z rozkładu normalnego  $N(\mu, \sigma)$  o **nieznanych**  $\mu$  i  $\sigma$ . Zmienna losowa

$$t = \frac{\bar{X} - \mu}{S_{n-1}} \sqrt{n}$$

ma rozkład  $t$  – *Studenta* o  $n - 1$  stopniach swobody  
(  $S_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$  ).

Po serii łatwych przekształceń :) otrzymujemy:

$$P(\bar{X} - t(\alpha/2, n-1)S_{n-1}/\sqrt{n} \leq \mu \leq \bar{X} + t(\alpha/2, n-1)S_{n-1}/\sqrt{n}) = 1 - \alpha.$$

Jest to przedział ufności dla parametru  $\mu$  utworzony na podstawie próby losowej  $X_1, X_2, \dots, X_n$  na poziomie istotności  $1 - \alpha$

Po serii łatwych przekształceń :) otrzymujemy:

$$P(\bar{X} - t(\alpha/2, n-1)S_{n-1}/\sqrt{n} \leq \mu \leq \bar{X} + t(\alpha/2, n-1)S_{n-1}/\sqrt{n}) = 1 - \alpha.$$

Jest to przedział ufności dla parametru  $\mu$  utworzony na podstawie próby losowej  $X_1, X_2, \dots, X_n$  na poziomie istotności  $1 - \alpha$

Uwaga

$t(\alpha/2, n - 1)$  oznacza kwantyl rzędu  $\alpha/2$  z rozkładu *studenta* z  $n - 1$  stopniami swobody.

Niech  $X_1, X_2, \dots, X_n$  będzie to próba losowa ( $n$ - „duże”) z rozkładu  $N(\mu, \sigma)$  przy czym parametr  $\mu$  oraz parametr  $\sigma$  jest **nieznany**.

### Uwaga

Przypadek ten traktujemy tak jak przypadek nr. 1 z tym, że we wzorze zamiast parametru  $\sigma$  stosujemy:

$$S_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

## Pytanie

Co zrobić gdy nasze obserwacje nie pochodzą z rozkładu normalnego?

## Możliwe odpowiedzi

- 1 w przypadku „małej” próby możemy napisać list do św. Mikołaja (w wiadomej sprawie);



## Pytanie

Co zrobić gdy nasze obserwacje nie pochodzą z rozkładu normalnego?

## Możliwe odpowiedzi

- 1 w przypadku „małej” próby możemy napisać list do św. Mikołaja (w wiadomej sprawie);
- 2 w przypadku „licznej” próby możemy użyć Centralnego Twierdzenia Granicznego.

## Wniosek z Centralnego Twierdzenia Granicznego

Jeśli z populacji o jakimkolwiek rozkładzie ze średnią  $\mu$  i odchyleniem standardowym  $\sigma$  pobieramy próby o dużej liczebności  $N$ , to rozkład średnich z tych prób będzie rozkładem normalnym o średniej  $\mu$  i odchyleniu standardowym  $\sigma/\sqrt{N}$ . Zatem dla dowolnych  $a, b$ ,  $a \leq b$  i zmiennej losowej  $Z$  o standardowym rozkładzie normalnym zachodzi

$$P\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) \approx P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$$

dla  $n$  dążących do nieskończoności, gdzie  $\Phi$  jest funkcją dystrybuanty standardowego rozkładu normalnego.

- 1 Nie ma uniwersalnej reguły mówiącej dla jak dużych  $n$  przybliżenie prawdopodobieństwa  $P(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b)$  przez rozkład normalny jest „dobre”

Rzucamy 200 razy nieznaną monetą. Podaj przedział ufności dla parametru  $p$  na poziomie ufności 0.95.

Stosujemy wzór:

$$P(\bar{X} - a_\alpha \sigma / \sqrt{n} \leq p \leq \bar{X} + a_\alpha \sigma / \sqrt{n}) = 1 - \alpha.$$

przy czym za  $\sigma$  wstawiamy  $\sqrt{\bar{X}(1 - \bar{X})}$ .

W celach antropometrycznych wylosowano  $n = 400$  studentów i dokonano pomiarów, mierząc między innymi długość ich stopy. Otrzymano z tej próby  $\bar{x} = 26.4$  oraz  $s = 1.7$  cm. Znajdź 0.90 przedział ufności dla średniej długości stopy.

W badaniach statystycznych wariancja należy do najczęściej szacowanych parametrów. Gdy obserwacje pochodzą z rozkładu normalnego, wtedy można zbudować przedział ufności dla wariancji  $\sigma^2$ .

Najczęściej używanymi estymatorami wariancji są statystyki określone wzorami:

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

oraz

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$



# Przypadek nr. 1

Niech  $X_1, X_2, \dots, X_n$  ( $n \leq 30$ ) będzie to próba losowa z rozkładu  $N(\mu, \sigma)$  przy czym parametr  $\mu$  oraz parametr  $\sigma$  jest **nieznany**. Po serii łatwych przekształceń :) otrzymujemy:

$$P\left(\frac{nS_n^2}{c_2} \leq \sigma^2 \leq \frac{nS_n^2}{c_1}\right) = 1 - \alpha.$$

gdzie  $c_1$  oraz  $c_2$  są wartościami kwantyli rozkładu  $\chi^2$  z  $n - 1$  stopniami swobody. Spełniającymi warunki:

$$P(\chi^2 < c_1) = \frac{1}{2}\alpha$$

oraz

$$P(\chi^2 \geq c_2) = \frac{1}{2}\alpha.$$

W celu oszacowania dokładności pewnego przyrządu pomiarowego dokonano nim 5 niezależnych pomiarów długości pewnego odcinka i otrzymano następujące wyniki:

15.15, 15.20, 15.04, 15.14, 15.22.

Przyjmując współczynnik ufności 0.98 zbudować przedział ufności dla nieznannej wariancji pomiarów tym przyrządem.