Statystyka w analizie i planowaniu eksperymentu

Paweł Błażej

28 marca 2012

Zdarzeniom losowym określonym na pewnej przestrzeni zdarzeń elementarnych Ω 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.

Przykład

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

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

oraz

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

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. $X_1, X_2, \ldots X_n$ (gdzie $X_1, X_2, \ldots$ (zmienne losowe)) nazywamy **próbą losową**

Jeżeli rozkłady zmiennych losowych $X_1, X_2, \ldots X_n$ 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.

Pojęcie próby prostej i statystyki

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

Przykłady statystyk

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. $mo,\ me,\ d,\ X_{min},\ X_{max}.$

Przykłady statystyk

Średnia z próby

Niech $X_1, X_2, \ldots X_n$ 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

Podstawowe własności średniej z próby

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

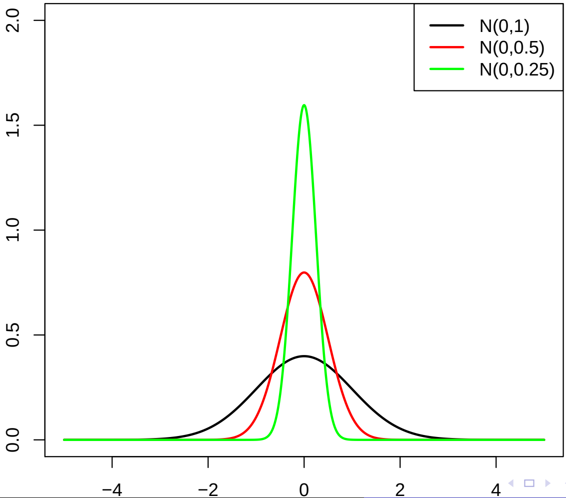
Podstawowe własności średniej z próby

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

Podstawowe własności średniej z próby

1. Jeżeli $EX_1 = \ldots = EX_n = \mu$, to $E\bar{X}_n = \mu$
2. Jeżeli $X_1, \ldots, X_n$ jest próbą prostą oraz $EX_i = \mu$ i $VarX_i = \sigma^2 < \infty$ dla $i = 1, 2, \ldots, n$, to $Var\bar{X}_n = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sigma^2$
3. Jeżeli $X_1, X_2, \ldots, X_n$ 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



Własności średniej - prawo wielkich liczb

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

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

Przykład zastosowania prawa wielkich liczb - rzut monetą

Nich $X_1, X_2, \ldots X_n$, 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.

Zadanie

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

Symulacje

```
[1] "n=50"
[1] 0.56
[1] "n=500"
[1] 0.518
[1] "n=5000"
[1] 0.5044
```

Uwaga

Prawo wielkich liczb nie daje żadnej odpowiedzi na pytanie jak liczna powinna być próba aby przybliżenie nieznanej 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 μ na podstawie próby $X_1, X_2, \ldots, X_n$ z rozkładu $\mathcal{N}(\mu, \sigma)$.

Problem

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

Uwaga

Rozpatrujemy dwa przypadki

1. parametr σ **jest znany** (rzadko spotykany w praktyce);
2. parametr σ **nie jest znany** (przypadek często spotykany w praktyce).

Przypadek nr. 1

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

Po serii łatwych przekształceń :) otrzymujemy:

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

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

Przypadek nr. 1

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

Po serii łatwych przekształceń :) otrzymujemy:

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

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

Uwaga

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

Przypadek nr. 2

Niech $X_1, X_2, \ldots, X_n$ będzie to próba losowa z rozkładu $\mathcal{N}(\mu, \sigma)$ przy czym parametr μ oraz parametr σ 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.

Przypadek nr. 2a

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

Przypadek 2a „szczypta”matematyki

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

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

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

Przypadek nr. 2a

Po serii łatwych przekształceń :) otrzymujemy:

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

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

Przypadek nr. 2a

Po serii łatwych przekształceń :) otrzymujemy:

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

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

Uwaga

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

Przypadek nr. 2b

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

Uwaga

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

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

Problem

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);

Problem

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.

Rozkład średnich z próby i błąd standardowy

Wniosek z Centralnego Twierdzenia Granicznego

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

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

dla n dążących do nieskończoności, gdzie Φ jest funkcją dystrybuanty standardowego rozkładu normalnego.

Uwaga

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

Zadanie

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

Odpowiedź

Stosujemy wzór:

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

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

Zadanie

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.

Przedział ufności dla wariancji

W badaniach statystycznych wariancja należy do najczęściej szacowanych parametrów. Gdy obserwacje pochodzą z rozkładu normalnego, wtedy można zbudować przedział ufności dla wariancji σ^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 $X_1, X_2, \ldots, X_n$ ($n \leqslant 30$) będzie to próba losowa z rozkładu $N(\mu, \sigma)$ przy czym parametr μ oraz parametr σ jest **nieznany**. Po serii łatwych przekształceń :) otrzymujemy:

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

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

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

oraz

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

Zadanie

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 nieznanej wariancji pomiarów tym przyrządem.