Rozkłady próbkowe

Rozważmy populację o pewnym rozkładzie, np.

- normalnym $N(\mu, \sigma)$ z parametrami μ i σ
- dwupunktowym z parametrem p
  (tzn. $P(Y = 1) = p$, $P(Y = 0) = 1 - p$)

Wybieramy próbę o rozmiarze n z populacji otrzymując:

- wartości $y_1, \ldots, y_n$
- łączną liczbę sukcesów y

Obliczamy estymatory:

- $\bar{y} = \frac{y_1 + \cdots + y_n}{n}$ oraz $s = \sqrt{\frac{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}}$
- $\hat{p} = \frac{y}{n}$

Gdy n jest duże, estymatory są na ogół bliskie parametrom.

Rozkłady próbkowe

- Jak bardzo estymatory mogą różnić się od prawdziwych parametrów?
- Co się stanie, gdy wylosujemy inną próbę?

Wyobraźmy sobie, że losujemy próbę wiele razy. Interesuje nas rozkład wszystkich możliwych do uzyskania wartości $\bar{y}$, s lub $\hat{p}$. Taki rozkład nazywamy *rozkładem próbkowym* estymatora.

Próbkowanie

Dana jest próba o rozmiarze n z populacji o rozkładzie normalnym. Obserwujemy średnią próbkową $\bar{y}$. Jaki jest rozkład tych średnich przy wielokrotnym próbkowaniu?

Fakt 1

Suma zmiennych niezeleżnych o rozkładzie normalnym ma rozkład normalny.

Fakt 2

Jeśli X ma rozkład normalny, to $aX + b$ dla $a \neq 0$ też ma rozkład normalny.

Wniosek

Jeśli Y ma rozkład normalny $N(\mu, \sigma)$, to rozkład próbkowy $\bar{Y}$ jest normalny $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Centralne Twierdzenie Graniczne

Centralne Twierdzenie Graniczne

Jeśli $Y_1, \ldots, Y_n$ to niezależne zmienne losowe o jednakowym rozkładzie i skończonej wariancji, to dla dużego n zmienna losowa $\bar{Y} = \frac{1}{n}(Y_1 + \ldots Y_n)$ ma rozkład bliski normalnemu.

- jeśli rozkład Y jest normalny, to dla dowolnego n rozkład $\bar{Y}$ jest normalny;
- jeśli rozkład Y jest w przybliżeniu symetryczny i nie ma „ciężkich ogonów”, to $n = 30$ jest wystarczająco duże.

Próbkowanie

Prosta próba losowa:

- dla każdego osobnika z populacji prawdopodobieństwo wybrania jest jednakowe,
- wybory poszczególnych osobników są od siebie niezależne.

Przykład 1

Amerykańska dziennikarka Ann Landers spytała swoich czytelników: *Gdybyście mogli zacząć życie ponownie, to czy znowu mielibyście dzieci?*. Odpisało prawie 10 000 czytelników i 70% odpowiedziało: NIE.

- odpowiadający to czytelnicy pisma Ann Landers, zaś populacja to wszyscy rodzice w USA;
- nawet wśród czytelników pisma, grupa osób odpowiadających nie była reprezentatywna (odpowiadają tylko ochotnicy);
- *Newsday* przeprowadził podobną ankietę na reprezentatywnej grupie 1373 rodziców i uzyskał 91% odpowiedzi: TAK.

Przykład 2

Magazyn *Literary Digest* w 1936 przeprowadził sondaż przedwyborczy wysłając kwestionariusz do 10 mln Amerykanów (25% głosujących). Odpowiedziało 2.4 mln osób.

- Prognozowany wynik: Landon 57%, Roosevelt 43%.
- Rzeczywisty wynik: Roosevelt 62%, Landon 38%.

- złe próbkowanie (brak reprezentatywności) – książki telefoniczne, członkostwo klubów, czytelnicy czasopisma itd.
- brak odpowiedzi (tylko 24% badanych odpowiedziało).

Obciążenie w próbkowaniu

Obciążenie w próbkowaniu

Obciążenie w próbkowaniu występuje, gdy mamy do czynienia z *systematycznym błędem* faworyzującym pewną część populacji. W przypadku takiego obciążenia nie pomoże duży rozmiar próby. Losowy wybór elemementów do próby zwykle eliminuje takie obciążenie.

Próbkowanie

Prosta próba losowa Dla każdego osobnika z populacji prawdopodobieństwo wyboru jest jednakowe i wybory poszczególnych osobników są od siebie niezależne.

Stratyfikacja Dzielimy populację na warstwy (podpopulacje podobnych jednostek) i oddzielnie próbkujemy w każdej warstwie, np. studenci i studentki.

Próbkowanie wielostopniowe Np. wybieramy losową próbę gmin, a w każdej wybranej gminie – losową próbę gospodarstw domowych.

Przedziały ufności

Przedział ufności

95% przedział ufności dla parametru populacji to przedział, dla którego mamy 95% pewności, że zawiera badany parametr.

Typowa sytuacja: nie znamy średniej populacji μ. Estymujemy ją przy pomocy średniej próbkowej $\bar{y}$. Kontruujemy taki przedział o środku $\bar{y}$, który z prawdopodobieństwem 95% zawiera μ.

Konstrukcja przedziału ufności

Załóżmy, że badana zmienna ma rozkład $N(\mu, \sigma)$, przy czym σ jest znane (założenie mało realistyczne, ale dobre na początek).
Wówczas średnia $\bar{Y}$ z n obserwacji ma rozkład $N(\mu, \frac{\sigma}{\sqrt{n}})$.
Dla standardowego rozkładu normalnego wyliczamy kwantyle rzędu 0.025 oraz 0.975:

$$P(Z > 1.96) = 0.025 \qquad P(Z < -1.96) = 0.025$$

Oznaczamy: $Z_{0.025} = 1.96$ Szukane kwantyle dla rozkładu $\bar{Y}$ wynoszą

$$\mu \pm Z_{0.025} \frac{\sigma}{\sqrt{n}}$$

Inaczej:

$$P(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Przykład 3

W USA przeprowadzono wśród studentów ankietę dotyczącą tego jaka kwota wydatków na studia pochodzi z funduszu oszczędnościowego (*college savings fund*). Spośród 1600 ankietowanych odpowiedziały 1593 osoby. Średnia odpowiedzi wyniosła \$1768. Zakładamy, że (skądś wiemy), że standardowe odchylenie dla całej populacji wynosi \$1483. Jaki jest 95% przedział ufności dla μ?

Rozkład nie jest normalny, ale z uwagi na dużą próbę, centralne twierdzenie graniczne gwarantuje nam, że przedział ufności wyliczany jak dla rozkładu normalnego będzie dobrym oszacowaniem.
Wyliczamy:

$$Z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\frac{1483}{\sqrt{1593}} \approx 73$$

Zatem mamy 95% pewności, że średnia kwot wydatków z funduszu oszczędnościowego to $\$1768 \pm 73$, czyli jest w przedziale $(1695, 1841)$.

Przykład 4

Jak w Przykładzie 3 średnia wynosi \$1768, a odchylenie standardowe dla całej populacji wynosi \$1483, ale tym razem liczba ankietowanych to 177 osób. Jaki jest 95% przedział ufności dla μ?

Wyliczamy:

$$Z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\frac{1483}{\sqrt{177}} \approx 218$$

Zatem mamy 95% pewności, że średnia kwot wydatków z funduszu oszczędnościowego to \$1768 $\pm$ 218, czyli jest w przedziale $(1550, 1986)$.

Przykład 5

Załóżmy, że planujemy ankietę jak w Przykładzie 3. Dopuszczalny margines błędu średniej kwoty to \$30 przy 95% poziomie ufności. Jaki rozmiar próby n potrzebujemy?

$$n = \left(\frac{1.96 \cdot 1483}{30} \right)^2 = 9387.54$$