
AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

# Estymacja przedziałowa

## Statystyka

**Dr inż. Janusz Majewski**
**Katedra Informatyki**

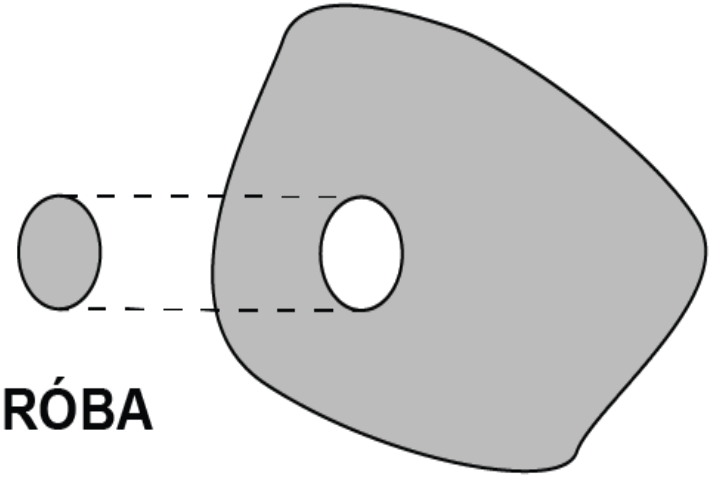
# Literatura

- Prezentacja wykorzystuje fragmenty książki: Amir D. Aczel „Statystyka w zarządzaniu”, PWN, 2007

# Populacja, próba, wnioskowanie

Statystyka jest nauką o wnioskowaniu, nauką o uogólnianiu polegającym na przechodzeniu od części (losowo wybranej próby) do całości (populacji). Populacja to zbiór wszystkich pomiarów, które nas interesują, a próba to podzbiór pomiarów wybranych z populacji. Losowa próba *n*-elementowa to próba wybrana w taki sposób, że każdy zbiór *n* elementów ma takie same szanse znalezienia sie w próbie, jak każdy inny zbiór *n* elementów populacji.

# Populacja, próba, estymacja



**PRÓBA**

**ZBIOROWOŚĆ GENERALNA (POPULACJA)**

na podstawie wyników próby → OBLICZAMY

CHARAKTERYZUJE SIĘ →

| | |
|---|---|
| średnia z próby | wartość oczekiwana (średnia) |
| wariancja z próby | wariancja populacji |
| odchylenie standardowe | odchylenie standardowe |

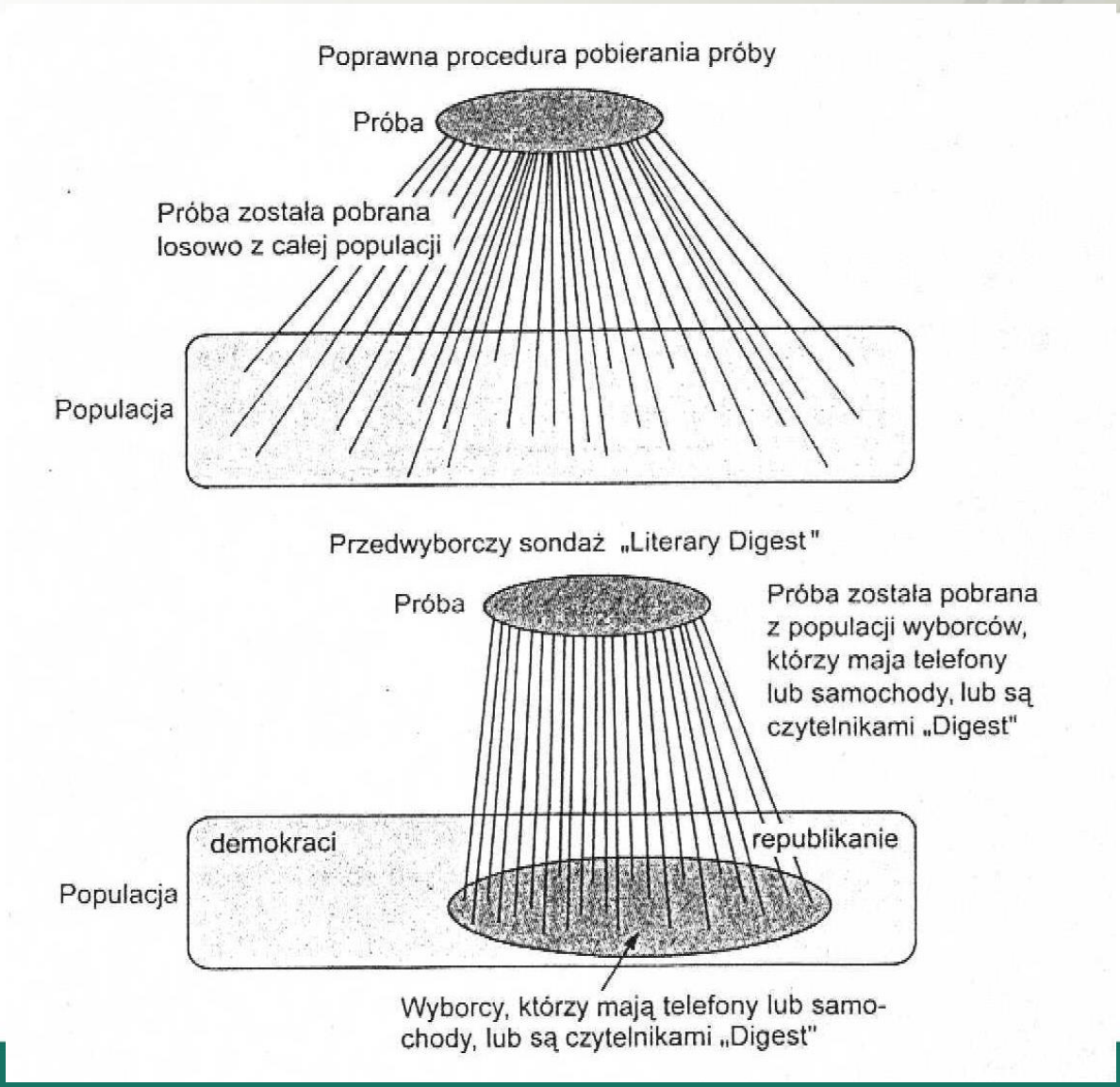
# Populacja, próba, estymacja

| | | |
|---|---|---|
| Średnia z próby: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ | jest estymatorem | Wartość średnia populacji: $\mu = \int_{-\infty}^{\infty} x\, f_x(x) dx$ |
| Wariancja z próby: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ | jest estymatorem | Wariancja populacji: $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\, f_x(x)\, dx$ |

$f_x(x)$ - jest nieznane

$\mu$ - jest nieznane

# Procedura pobierania próby



# Wybory prezydenckie w USA w 1936 r.

**Sondaż** przedwyborczy czasopisma „Literary Digest”:

- Gubernator Kansas Alfred M. Landon (republikanin) zwycięży w 32 stanach na 48 stanów i pokona prezydenta Franklina D. Roosevelta (demokratę).

**Wyniki wyborów**:

- Franklin D. Roosevelt wygrał w 46 stanach i został wybrany prezydentem USA pokonując Alfreda M. Landona, który zwyciężył zaledwie w 2 stanach.

# Estymacja

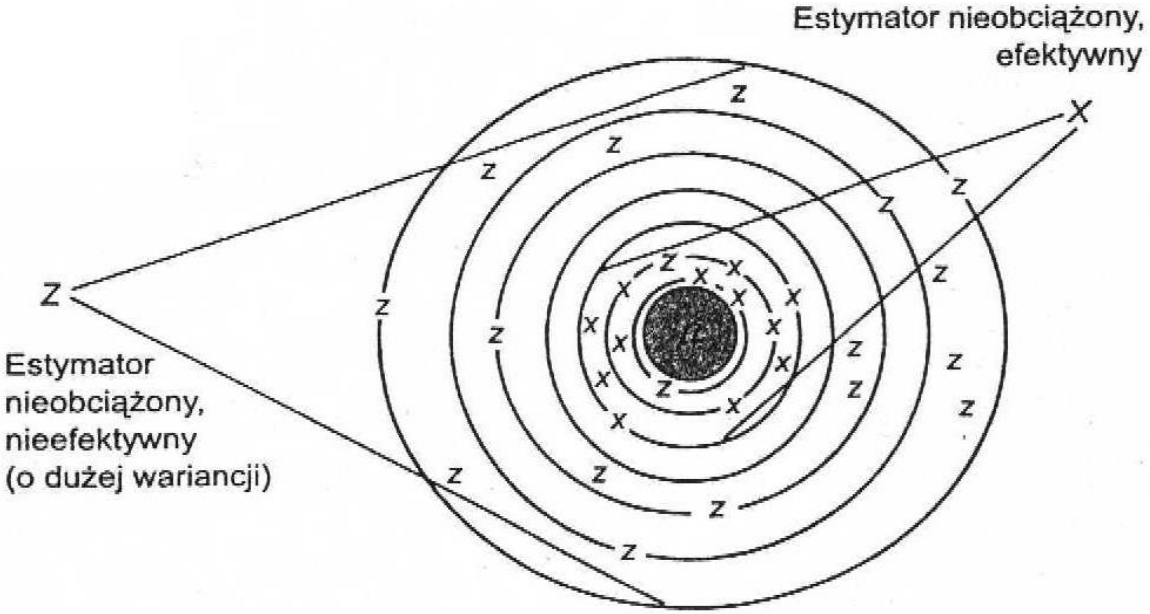
***Statystyka z próby***: zmienna losowa będąca dowolną funkcją wyników próby losowej, np. średnia z próby $\bar{x}$, wariancja z próby $s^2$

***Estymator:*** dowolna statystyka $Z$ służąca do oszacowania nieznanej wartośći $\theta$ parametru populacji generalnej.

***Estymator efektywny:*** estymator $Z$ o możliwie małym rozrzucie (małej wariancji $\sigma^2(Z)$). Stosowanie estymatora efektywnego oznacza popełnianie małego błędu przeciętnego szacunku.

# Estymator efektywny

Estymator jest **efektywny**, jeżeli ma niewielką wariancję (a tym samym niewielkie odchylenie standardowe).




**Rysunek 5.11.** Dwa nieobciążone estymatory parametru $\mu$; estymator $X$ jest efektywniejszy od estymatora $Z$

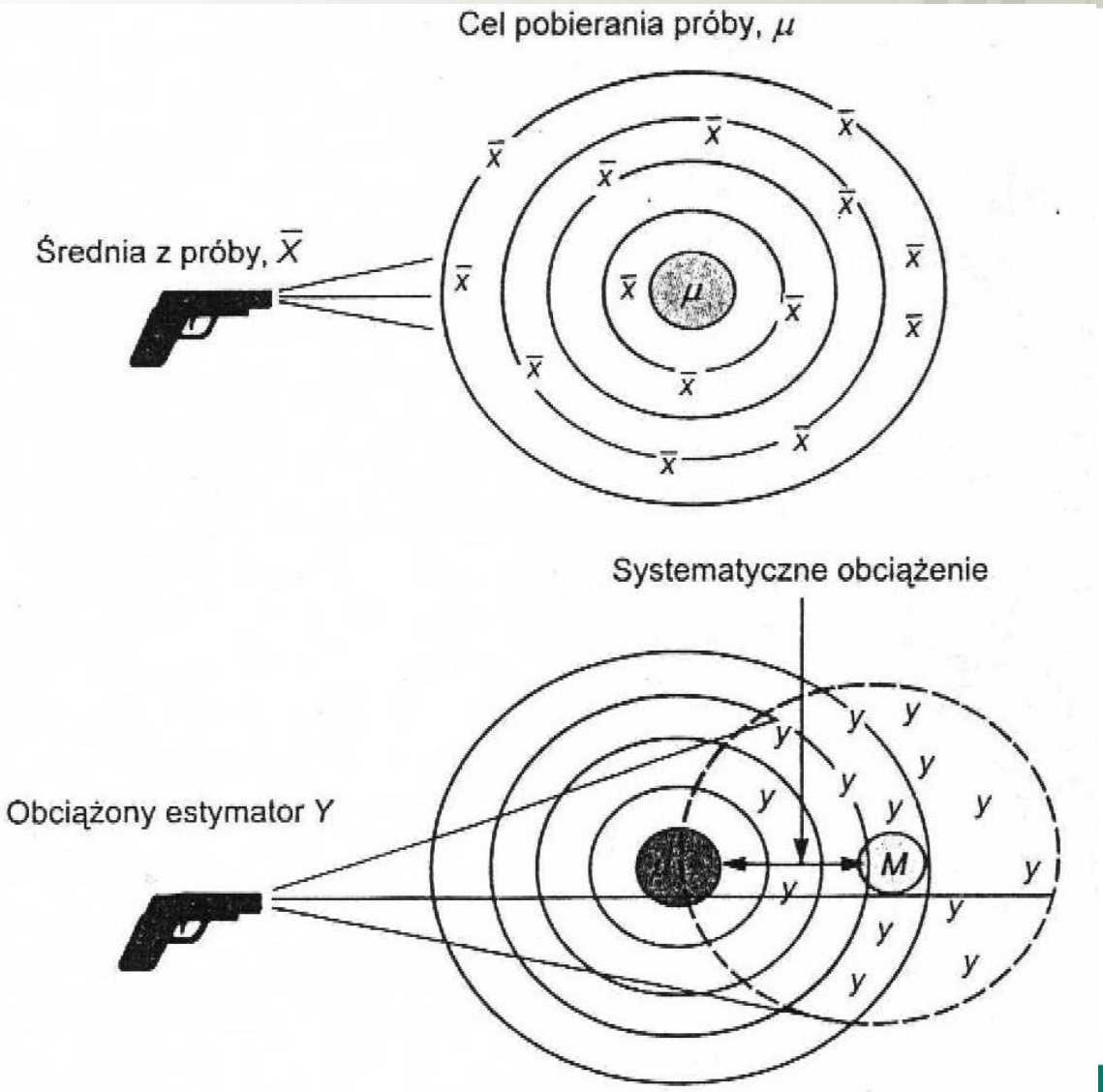
# Estymacja

***Estymator zgodny:*** estymator $Z$ parametru $\theta$ spełniający warunek $\lim_{n\to\infty} P\{|Z_n - \theta| < \varepsilon\} = 1$, tzn. estymator zapewniający, że stosowanie większych liczebności próby poprawia dokładności szacunku.

***Estymator dostateczny:*** estymator $Z$ wykorzystujący wszystkie informacje (wszystkie dane) zawarte w próbie (np. mediana z próby nie jest estymatorem dostatecznym, gdyż zależy tylko od jednej lub dwóch „środkowych”danych z próby).

***Estymator nieobciążony:*** estymator $Z$ spełniający równość $E(Z) = \theta$, oznaczającą, że $Z$ szacuje nieznaną wartość $\theta$ parametru populacji bez błędu systematycznego.

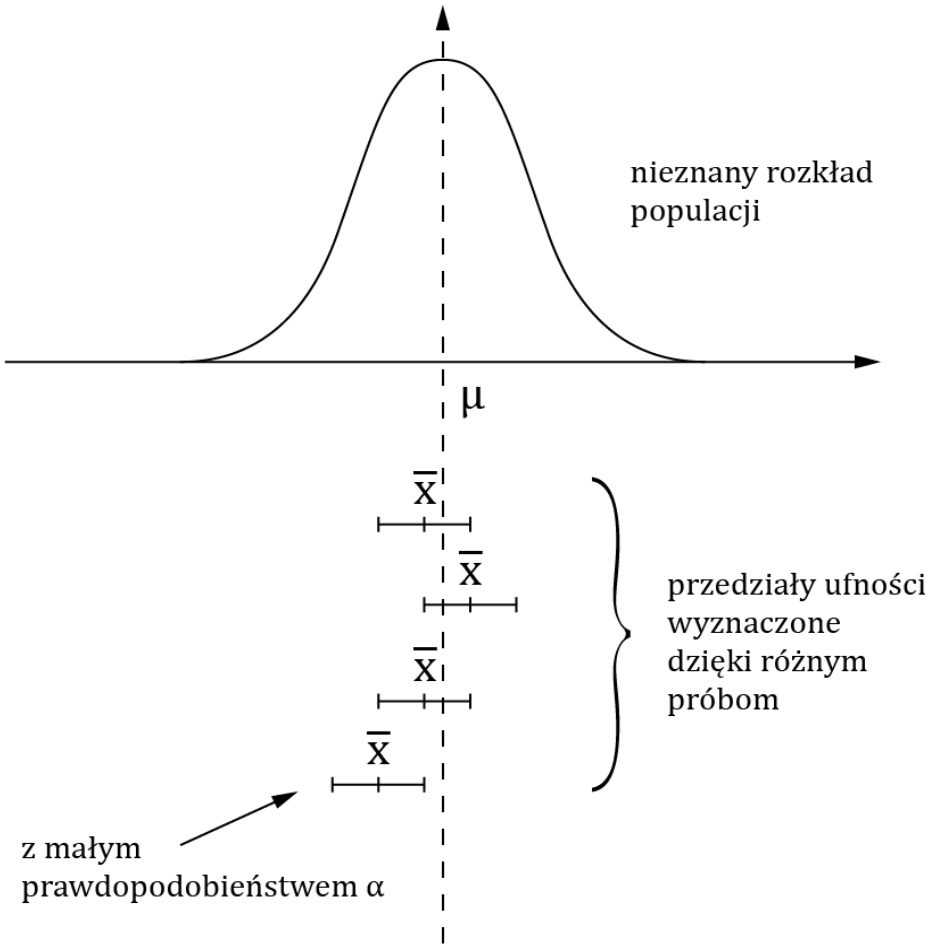
# Estymator nieobciążony



# Estymacja przedziałowa

Estymacja przedziałowa polega na szacowaniu nieznanego parametru populacji poprzez budowę przedziału, który z zadanym z góry prawdopodobieństwem pokrywałby nieznaną wartość parametru. Poszukiwany w zadaniu przedział nazywany jest *przedziałem ufności*, zaś ustalane a priori prawdopodobieństwo, z którym przedział ufności ma pokrywać nieznaną wartość parametru, nosi nazwę *współczynnika ufności* (*poziomu ufności*).

# Estymacja przedziałowa



# Estymacja przedziałowa

**Table 1a. Age-adjusted prevalence of diagnosed and undiagnosed diabetes among adults aged ≥18 years, United States, 2011–2014**

| Characteristic | Diagnosed diabetes Percentage (95% CI) | Undiagnosed diabetes Percentage (95% CI) | Total Percentage (95% CI) |
|---|---|---|---|
| **Total** | **8.7 (8.1–9.4)** | **2.7 (2.3–3.3)** | **11.5 (10.7–12.4)** |
| **Sex** | | | |
| Women | 8.5 (7.5–9.5) | 2.3 (1.8–3.1) | 10.8 (9.8–11.9) |
| Men | 9.1 (8.4–9.9) | 3.2 (2.4–4.3) | 12.3 (11.3–13.4) |
| **Race/Ethnicity** | | | |
| Asian, non-Hispanic | 10.3 (8.6–12.4) | 5.7 (4.0–8.2) | 16.0 (13.6–18.9) |
| Black, non-Hispanic | 13.4 (12.2–14.6) | 4.4 (3.0–6.2) | 17.7 (15.8–19.9) |
| Hispanic | 11.9 (10.3–13.7) | 4.5 (3.2–6.2) | 16.4 (14.1–18.9) |
| White, non-Hispanic | 7.3 (6.6–8.1) | 2.0 (1.5–2.6) | 9.3 (8.4–10.2) |
| **Education** | | | |
| Less than high school | 11.4 (9.9–13.1) | 4.1 (3.0–5.6) | 15.5 (13.5–17.7) |
| High school | 10.3 (8.8–12.0) | 3.2 (2.4–4.2) | 13.5 (11.9–15.2) |
| More than high school | 7.4 (6.6–8.4) | 2.2 (1.6–3.0) | 9.6 (8.6–10.7) |

CI = confidence interval.

Data source: 2011–2014 National Health and Nutrition Examination Survey.

https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

# Estymacja przedziałowa średniej

Szacowanym nieznanym parametrem jest nieznana średnia populacji $\mu$. Szacowanie odbywa się na podstawie średniej $\bar{x}$ z próby zawierającej $n$ elementów. Średnia z próby $\bar{x}$ jest zmienną losową: o własnościach wyspecyfikowanych na następnym slajdzie.

# Estymacja przedziałowa średniej
## (centralne twierdzenie graniczne)

a) $E(\bar{x}) = \mu$

Wszystkie średnie z prób grupują się wokół rzeczywistej średniej populacji

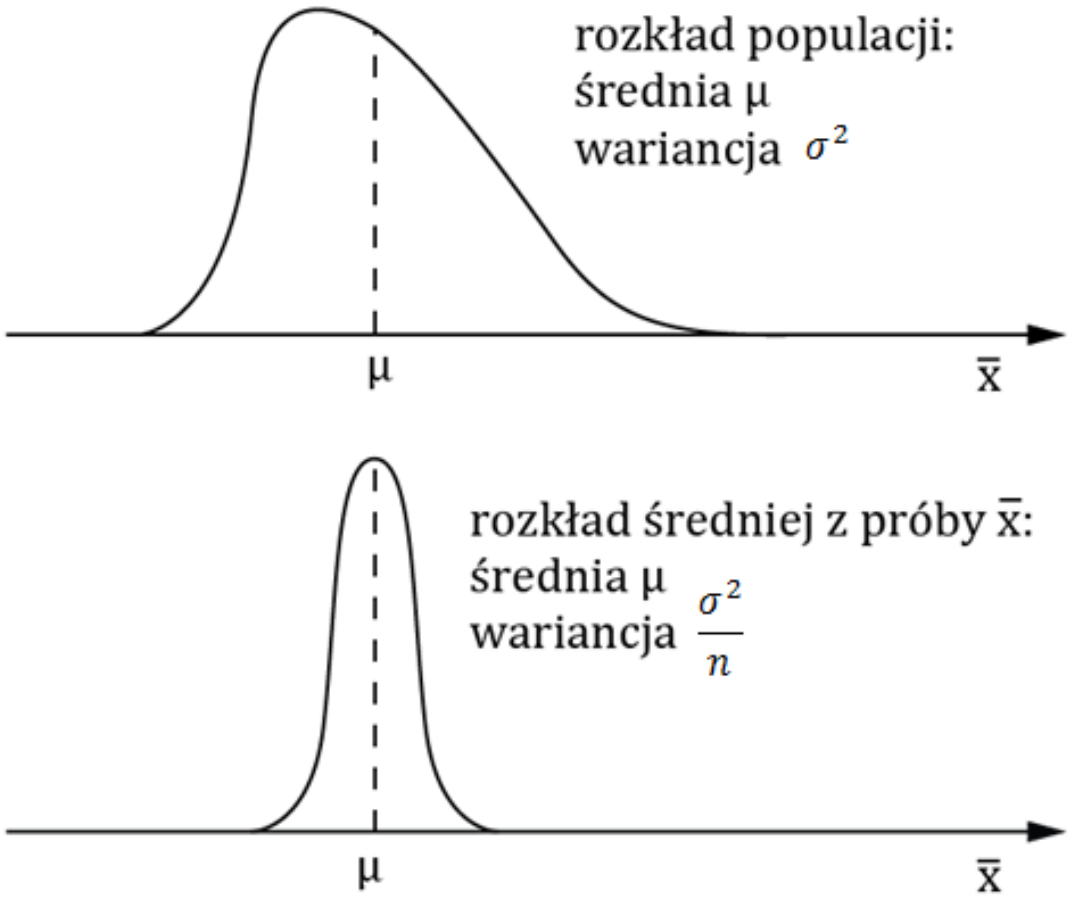
b) $\sigma^2(\bar{x}) = \frac{\sigma^2}{n}$

$\boldsymbol{\sigma^2}$ - prawdziwa wariancja populacji, $\boldsymbol{n}$ - liczebność próby

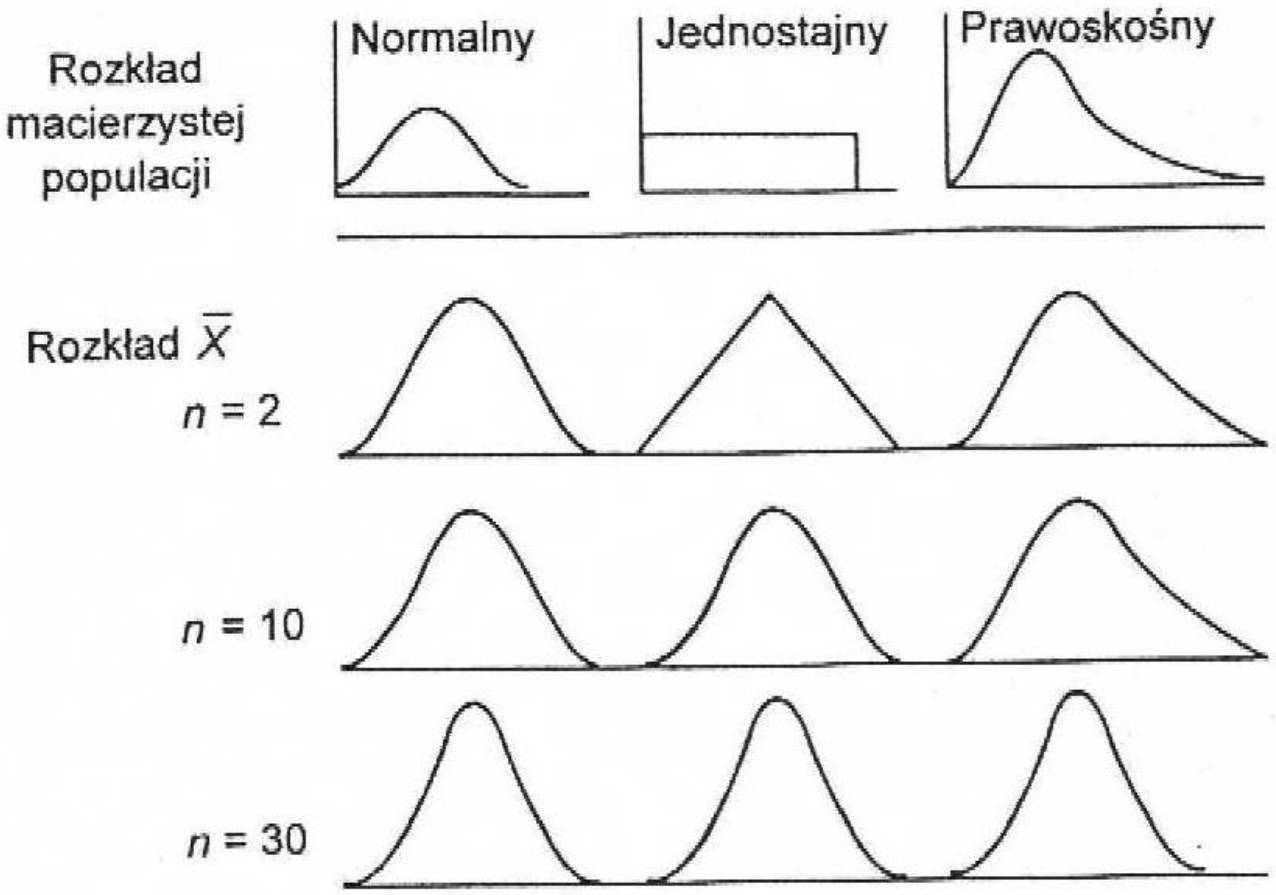
Średnie z prób mają tym mniejszy rozrzut im większa jest próba, a tym większy rozrzut, im większy jest rozrzut w populacji generalnej.

c) Jeżeli rozkład populacji jest normalny, to rozkład średniej z próby też jest normalny.
Jeżeli nawet rozkład populacji odbiega od normalnego, to rozkład średniej z próby w miarę wzrostu $n$ zbliża się do rozkładu normalnego

# Estymacja przedziałowa średniej

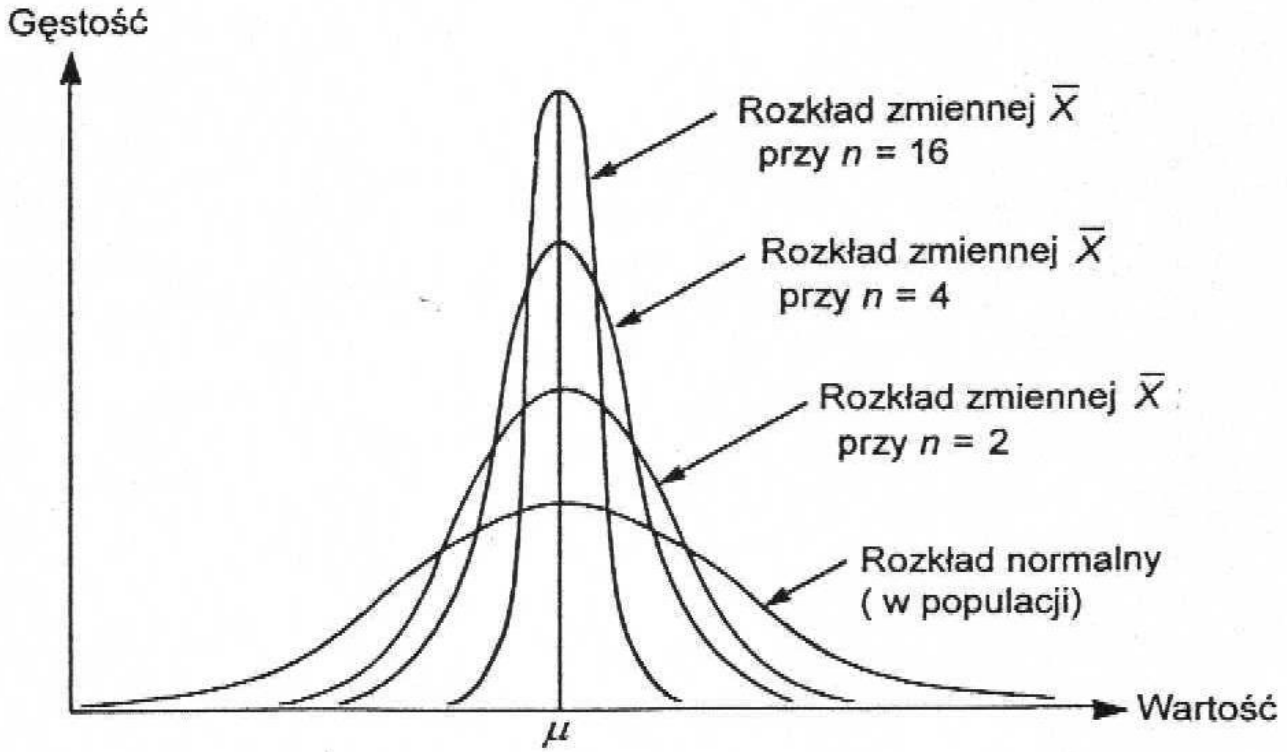


# Rozkład średniej z próby




**Rysunek 5.6.** Znaczenie centralnego twierdzenia granicznego: rozkład średniej z próby, $\overline{X}$, dla różnych populacji i różnych liczebności próby

# Rozkład średniej z próby




**Rysunek 5.4.** Normalny rozkład w populacji i rozkład średniej z próby, $\overline{X}$, o różnej liczebności

# Estymacja przedziałowa średniej

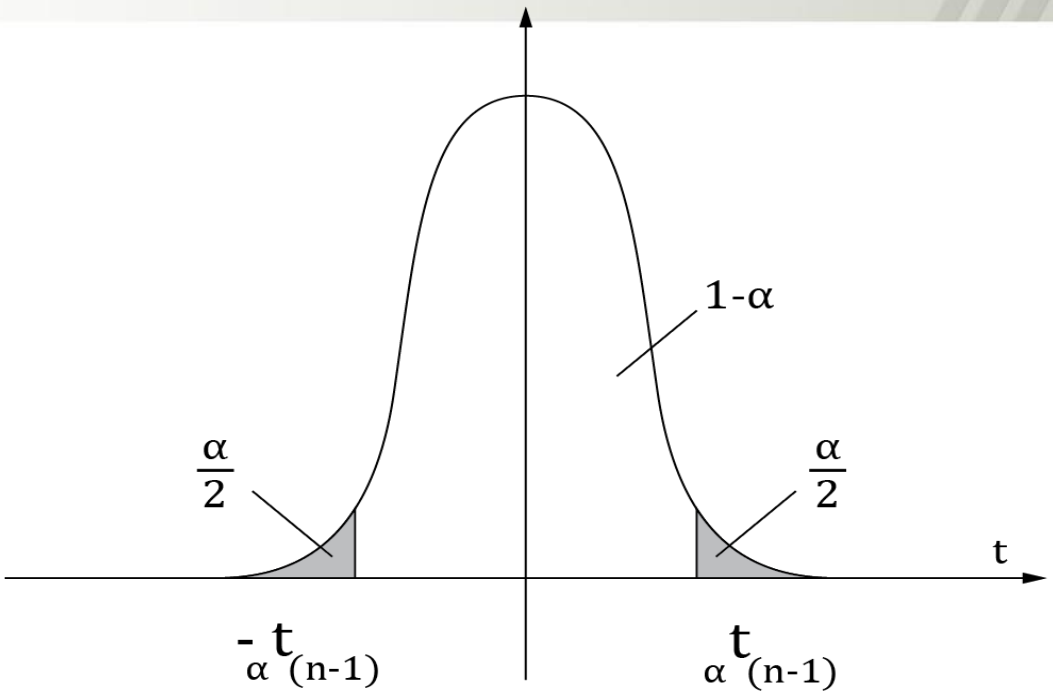
Z faktu, że $\bar{x}$ ma rozkład normalny lub zbliżony do normalnego nie można bezpośrednio skorzystać, bo $\sigma^2$ populacji nie jest znana. Wykorzystujemy statystykę t:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\frac{s}{\sqrt{n}}$ - Odchylenie standardowe średniej z próby zwane *błędem standardowym średniej*

Statystyka $t$ ma rozkład t-Studenta o $n - 1$ stopniach swobody

# Estymacja przedziałowa średniej




Z dużym prawdopodobieństwem $1-\alpha$ zmienna t wyznaczona z próby znajduje się w przedziale:

$$-{}_{\alpha}t_{(n-1)} < t < {}_{\alpha}t_{(n-1)}$$

${}_{\alpha}t_{(n-1)}$ - wart. krytyczna odczytana z tablic

# Estymacja przedziałowa średniej

$$-_{\alpha}t_{(n-1)} < t < {}_{\alpha}t_{(n-1)}$$

uwzględniając, że

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

mamy

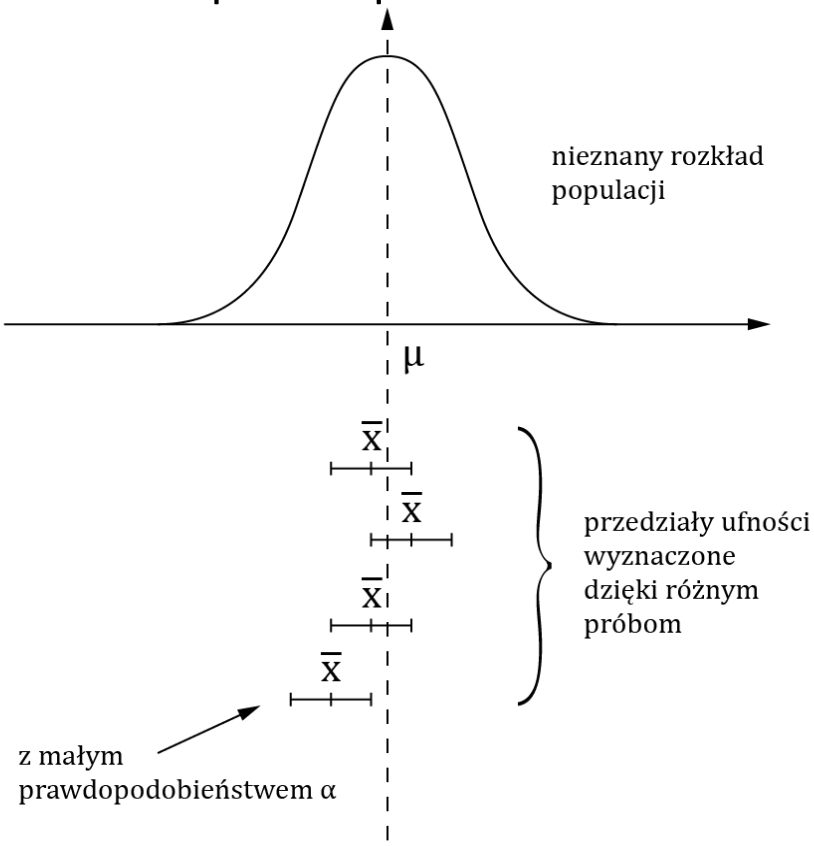
$$-_{\alpha}t_{(n-1)} < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < {}_{\alpha}t_{(n-1)}$$

$$\bar{x} - \frac{{}_{\alpha}t_{(n-1)}\, s}{\sqrt{n}} < \mu < \bar{x} + \frac{{}_{\alpha}t_{(n-1)}\, s}{\sqrt{n}}$$

# Estymacja przedziałowa średniej

$$\bar{x} - \frac{\alpha t_{(n-1)}\, s}{\sqrt{n}} < \mu < \bar{x} + \frac{\alpha t_{(n-1)}\, s}{\sqrt{n}}$$

Tak określony przedział ufności z prawdopodobieństwem $1 - \alpha$ *obejmuje* nieznaną wartość $\mu$.



# Estymacja przedziałowa średniej

Na ogół poziom ufności $1-\alpha$ wynosi **0.9, 0.95, 0.99** .

Im bliższy 1 jest współczynnik ufności, tym szerszy przedział ufności.

Dla dużych prób ($n > 100$) zamiast statystyki $t$ stosuje się statystykę $u$:

$$u = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

mającą standaryzowany rozkład normalny.

Wówczas przedział ufności dla średniej jest określany jako:

$$\bar{x} - \frac{{}_{\alpha}u\, s}{\sqrt{n}} < \mu < \bar{x} + \frac{{}_{\alpha}u\, s}{\sqrt{n}}$$

# Estymacja przedziałowa średniej

Wartości krytyczne ${}_{\alpha}u$ rozkładu normalnego standaryzowanego

| $1-\alpha$ | $\alpha$ | ${}_{\alpha}u$ |
|---|---|---|
| 0.9 | 0.1 | 1.645 |
| **0.95** | **0.05** | **1.960** |
| 0.99 | 0.01 | 2.576 |

# Estymacja przedziałowa średniej

Wiek pacjentek z nowotworem szyjki macicy w pewnym szpitalu w Algierii

| wiek | liczba pacjentek | środek przedziału wiekowego | D = ((C)-50.9)**2 | E = (D) * (B) |
|---|---|---|---|---|
| (A) | (B) | (C) | (D) | (E) |
| 20-25 | 3 | 22,5 | 806,6 | 2419,68 |
| 25-30 | 10 | 25,5 | 645,2 | 6451,6 |
| 30-35 | 38 | 32,5 | 338,6 | 12865,28 |
| 35-40 | 71 | 37,5 | 179,6 | 12748,76 |
| 40-45 | 117 | 42,5 | 70,6 | 8255,52 |
| 45-50 | 100 | 47,5 | 11,6 | 1156 |
| 50-55 | 89 | 52,5 | 2,6 | 227,84 |
| 55-60 | 75 | 57,5 | 43,6 | 3267 |
| 60-65 | 70 | 62,5 | 134,6 | 9419,2 |
| 65-70 | 59 | 67,5 | 275,6 | 16258,04 |
| 70-75 | 21 | 72,5 | 466,6 | 9797,76 |
| 75-80 | 11 | 77,5 | 707,6 | 7783,16 |
| 80-85 | 1 | 82,5 | 998,6 | 998,56 |
| 85-90 | 2 | 87,5 | 1339,6 | 2679,12 |
| Suma | 667 | | | 94327,52 |

| | |
|---|---|
| Średnia z próby: | 50,9 |
| Wariancja z próby: | 141,6 |
| Odchylenie stand. z próby: | 11,9 |
| Błąd stand. średniej: | 0,5 |
| **Szer. 99% przedz. ufności śr.:** | **1,2** |
| Dolna granica przedz. ufności.: | 49,7 |
| Górna granica przedz. ufności.: | 52,1 |
| **Szer. 95% przedz. ufności śr.:** | **0,9** |
| Dolna granica przedz. ufności.: | 50 |
| Górna granica przedz. ufności.: | 51,8 |

# Estymacja przedziałowa frakcji

Próba polegała na wykonaniu $n$ doświadczeń, z których $r$ dało wynik pozytywny (sukces).

Frakcja obliczona z próby:

$$p = \frac{r}{n}$$

jest estymatorem nieznanego parametru $\pi$ (prawdopodobieństwo sukcesu) populacji generalnej, przy czym:

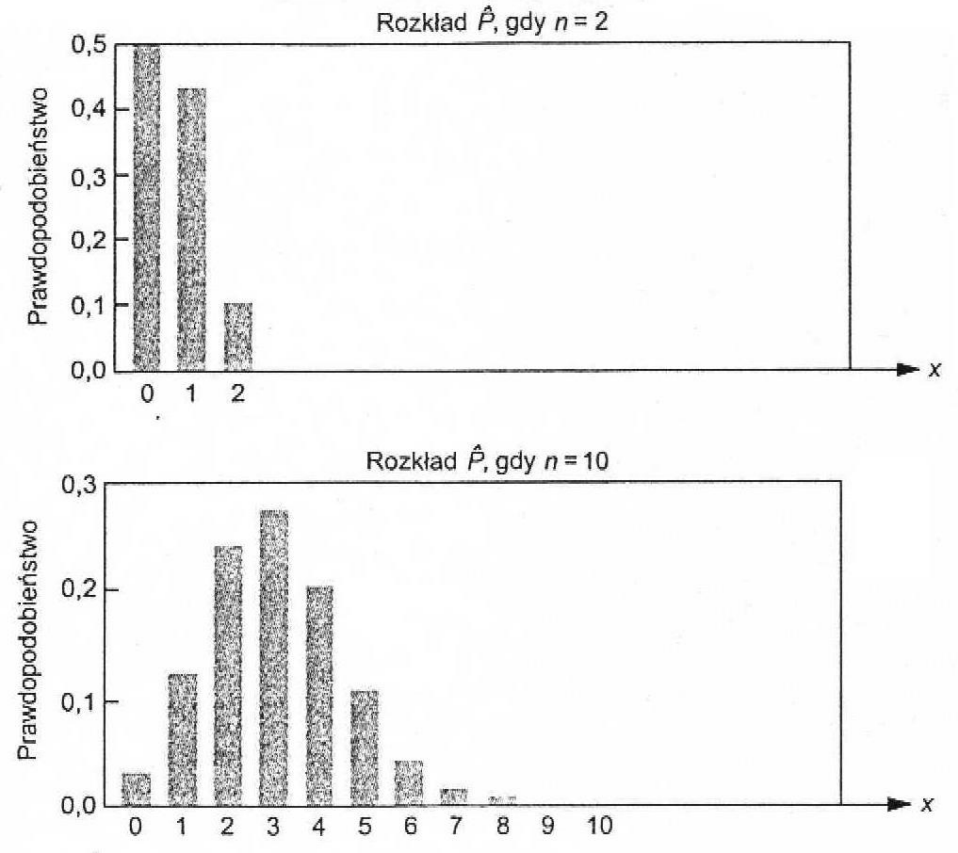
$$E(p) = \pi$$

$$\sigma^2(p) = \frac{\pi\,(1 - \pi)}{n}$$

zaś zmienna $r = p\,n$ ma rozkład dwumianowy.

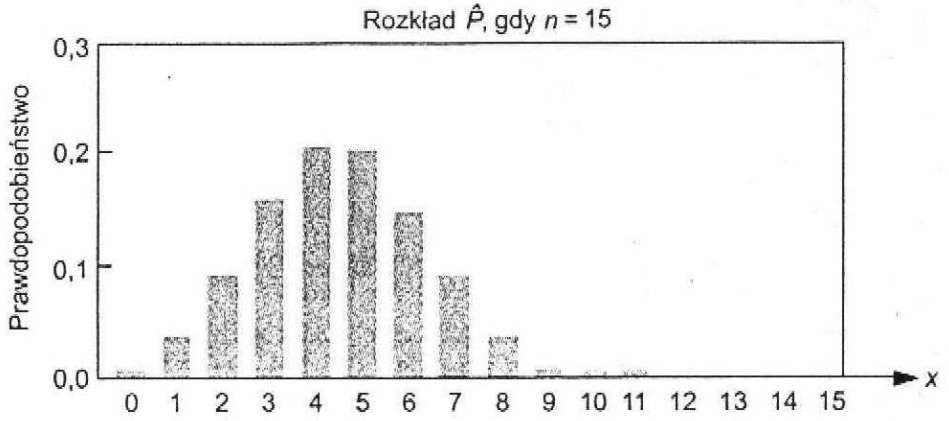
Rozkład ten w miarę wzrostu wielkości próby $n$ zmierza do rozkładu normalnego.

# Rozkład frakcji z próby









# Estymacja przedziałowa frakcji

Załóżmy, że próba jest duża ($n > 100$) oraz nieznany parametr $\pi$ nie jest zbyt mały ($\pi > 0.05$ ). Wówczas przedział ufności dla $\pi$ ze współczynnikiem ufności równym $1 - \alpha$ jest określony jako:

$$p \pm_{\alpha} u \sqrt{\frac{p(1-p)}{n}}$$

$\sigma(p) = \sqrt{\frac{p(1-p)}{n}}$ - Błąd standardowy frakcji

Jeśli założenia powyższe nie są spełnione, należy wykorzystać informacje o rozkładzie dwumianowym (patrz [Parker]).

# Estymacja przedziałowa frakcji

**Przykład:** Wylosowano 150 studentów pewnej uczelni, z nich 105 paliło tytoń.

$$n = 150 \qquad r = 105 \qquad p = \frac{105}{150} = 0.7$$

$$_{0.05}u = 1.96$$

$$_{\alpha}u\sqrt{\frac{p(1-p)}{n}} = 0.073 \approx 0.07$$

$$\pi = 0.7 \pm 0.07$$

$$0.63 < \pi < 0.77$$

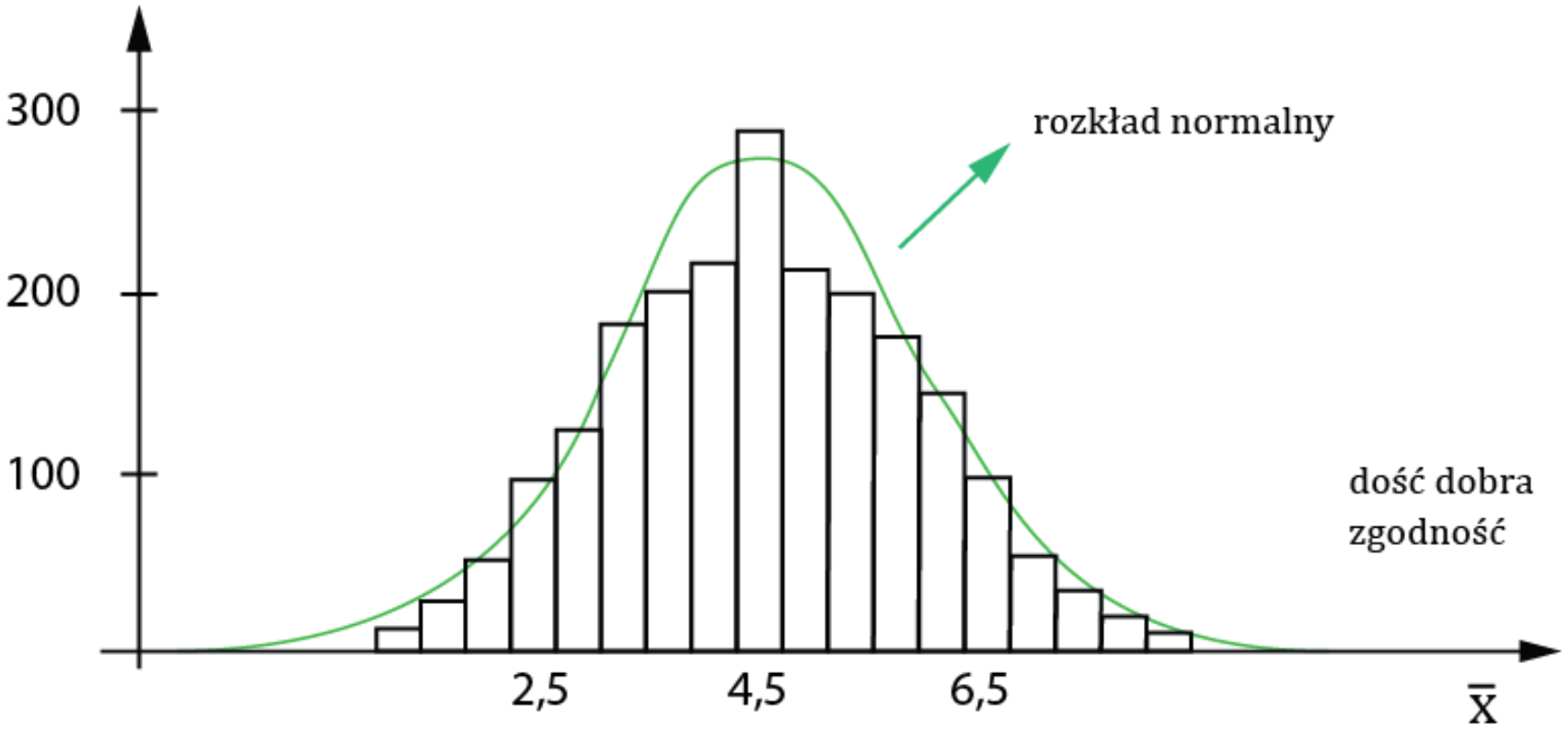
# Estymacja przedziałowa wariancji

Wszelkie sposoby wnioskowania dotyczące wariancji są znacznie bardziej czułe na odchyłki rozkładu rzeczywistego populacji od rozkładu normalnego, niż metody dotyczące średniej. Należy więc najpierw sprawdzić, czy można uważać, iż próbka została wylosowana z populacji o rozkładzie normalnym.

**Przykład:** Wylosowano 2000 próbek 5-elementowych w populacji o **rozkładzie skokowym równomiernym** (10 cyfr od 0 do 9 z prawdopodobieństwem 1/10 dla każdej cyfry). Uzyskano 2000 średnich $\bar{x}$ z tych próbek oraz 2000 oszacowań wariancji $s^2$.Oto histogramy średnich i wariancji wraz z rozkładami teoretycznymi.

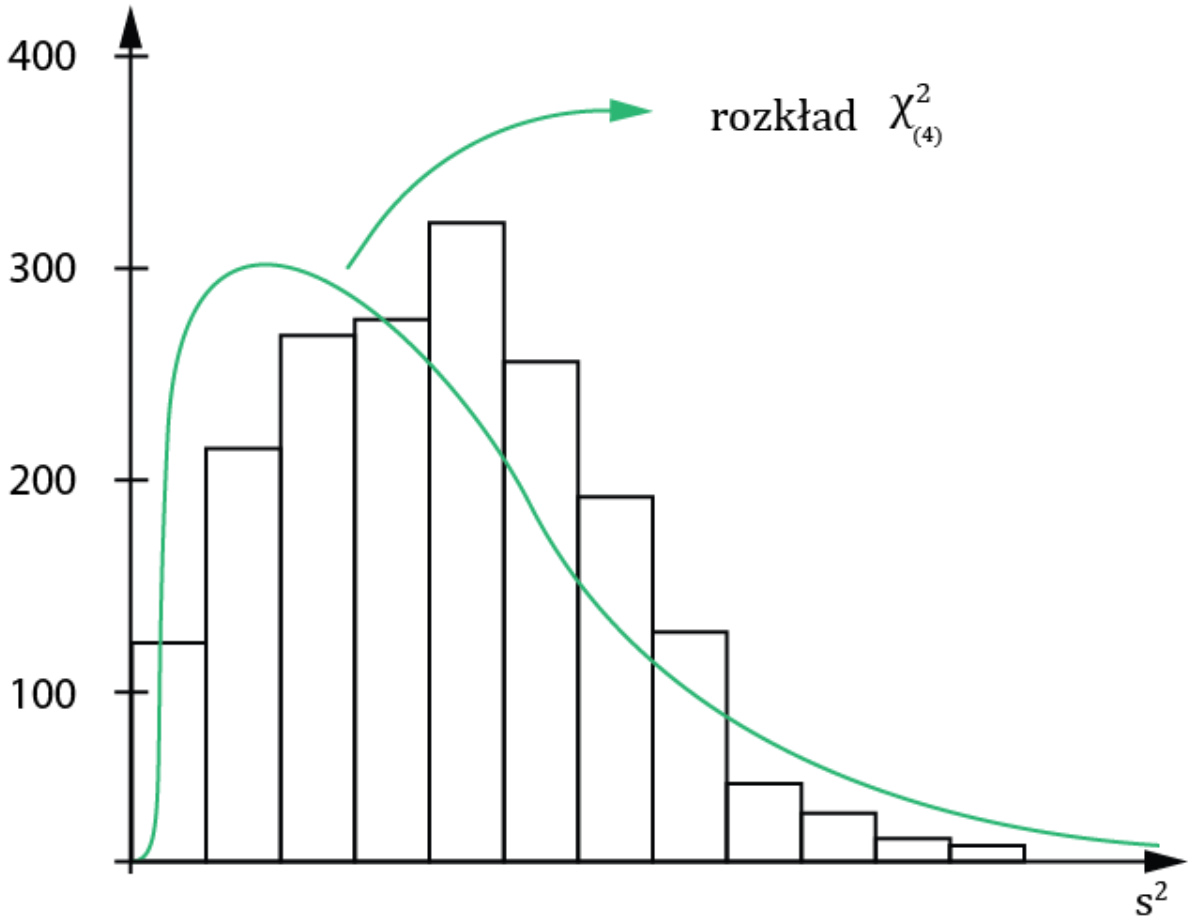
# Estymacja przedziałowa wariancji

HISTOGRAM ŚREDNICH Z PRÓBY



# Estymacja przedziałowa wariancji

HISTOGRAM WARIANCJI Z PRÓBY



# Estymacja przedziałowa wariancji

Estymatorem nieznanej wariancji populacji $\sigma^2$ jest nieobciążony estymator $s^2$, przy czym:

$$E(s^2) = \sigma^2$$

$$\sigma^2(s^2) = \frac{\sigma^2}{n-1}$$

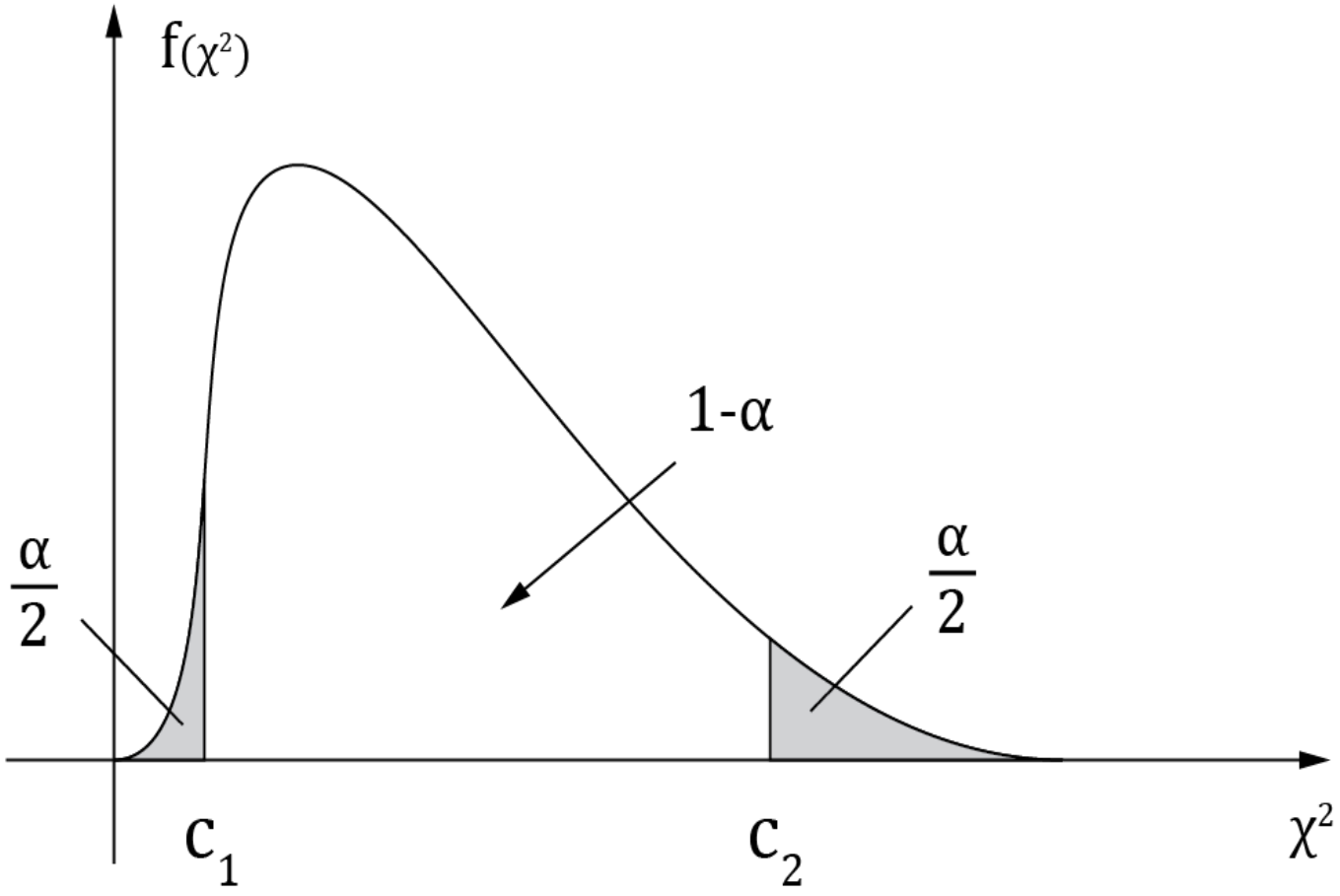
gdzie: $n$ - liczebność próby

Statystyka $\chi^2$:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

ma rozkład $\chi^2$ o $n-1$ stopniach swobody

# Estymacja przedziałowa wariancji



# Estymacja przedziałowa wariancji

Przedział ufności dla $\sigma^2$ ze współczynnikiem ufności $1-\alpha$ określamy jako:

$$\frac{(n-1)s^2}{c_2} < \sigma^2 < \frac{(n-1)s^2}{c_1}$$

$c_1 = {}_{1-\frac{\alpha}{2}}\chi^2_{(n-1)}$ oraz $c_2 = {}_{\frac{\alpha}{2}}\chi^2_{(n-1)}$ - wartości krytyczne odczytane z tablic

# Estymacja przedziałowa wariancji

Jeżeli dysponujemy dużą próbą $(n > 100)$ pobraną z populacji o rozkładzie normalnym lub zbliżonym do normalnego, to na podstawie estymatora $s$ wyznaczonego z tej próby możemy w przybliżeniu oszacować odchylenie standardowe $\sigma$ populacji według poniższej zależności:

$$\frac{s}{1 + \frac{{}_{\alpha}u}{\sqrt{2n}}} < \sigma < \frac{s}{1 - \frac{{}_{\alpha}u}{\sqrt{2n}}}$$

gdzie ${}_{\alpha}u$ jest wartością krytyczną standaryzowanego rozkładu normalnego.

# Szacowanie niezbędnej liczebności próby

Szacowanie niezbędnej liczby doświadczeń przy estymowaniu średniej

Wykonaliśmy $n_0$ pomiarów. Na podstawie $n_0$-elementowej próby próbujemy oszacować jaka jest niezbędna liczba pomiarów, aby uzyskać z góry założoną szerokość $d$ przedziału ufności dla średniej.

$\bar{x} \pm d$ - przedział ufności

$d = {}_{\alpha}t\frac{s}{\sqrt{n}}$ stąd $n = \frac{s^2\,{}_{\alpha}t^2}{d^2}$

$s^2$, ${}_{\alpha}t$ - na podstawie $n_0$ pomiarów

Należy jeszcze wykonać około $n - n_0$ pomiarów

# Szacowanie niezbędnej liczebności próby

**Przykład:** W badaniu taksonomicznym chcemy oszacować długość określonej struktury z danej populacji z dokładnością $\pm 1.00$ mm przy poziomie ufności 0.95. Po wykonaniu 26 pomiarów oszacowano odchylenie standardowe, które wynosiło 4 mm. Ile pomiarów należy jeszcze wykonać?

$d = 1, \qquad s = 4, \qquad {}_{0.05}t_{(25)} = 2.06, \qquad n_0 = 26$

$$n = \frac{(2.06)^2\ 4^2}{1^2} = 69.32\ \approx 70$$

Należy jeszcze wykonać około $n - n_0 = 70 - 26 = 44$ pomiary.

# Szacowanie niezbędnej liczebności próby

Szacowanie niezbędnej liczby doświadczeń przy estymowaniu frakcji

$p \pm d$ - przedział ufności

$$d = {}_{\alpha}u \sqrt{\frac{p\,(1-p)}{n}}$$

$$n = \frac{{}_{\alpha}u^2\; p\,(1-p)}{d^2}$$

$p\,(1-p)$ - określane na podstawie $n_0$ doświadczeń

Należy wykonać dodatkowo $n - n_0$ doświadczeń.

# Szacowanie niezbędnej liczebności próby

**Przykład:** Przeprowadzono wywiad z 150 studentami. Z tego 105 paliło systematycznie tytoń. Z iloma studentami należy dodatkowo przeprowadzić wywiad, aby oszacować procent palących z dokładnością co najwyżej $\pm 3\%$ przy współczynniku ufności $1 - \alpha = 0.95$

$$n_0 = 150 \qquad p = \frac{105}{150} = 0.7 \qquad {}_{\alpha}u = 1.96 \qquad d = 0.03$$

$$n = \frac{(1.96)^2 \cdot 0.7 \cdot 0.3}{(0.03)^2} = 896.4 \approx 897$$

Należy jeszcze dodatkowo rozmawiać z około $n - n_0 = 897 - 150 = 747$ studentami.

# Szacowanie niezbędnej liczebności próby

**Przykład c.d. (to samo zadanie dla innych wymagań):** Przeprowadzono wywiad z 150 studentami. Z tego 105 paliło systematycznie tytoń. Z iloma studentami należy dodatkowo przeprowadzić wywiad, aby oszacować procent palących z dokładnością co najwyżej $\pm 2\%$ przy współczynniku ufności $1 - \alpha = 0.95$

$$n_0 = 150 \qquad p = \frac{105}{150} = 0.7 \qquad {}_{\alpha}u = 1.96 \qquad d = 0.02$$

$$n = \frac{(1.96)^2 \cdot 0.7 \cdot 0.3}{(0.02)^2} = 2016.8 \approx \mathbf{2017}$$

Należy jeszcze dodatkowo rozmawiać z około $n - n_0 = 2017 - 150 = 1867$ studentami.

# Szacowanie niezbędnej liczebności próby

**Przykład c.d. (jeszcze raz to samo zadanie, ale dla innych danych):**
Przeprowadzono wywiad z 150 studentami. Z tego **75** paliło systematycznie tytoń. Z iloma studentami należy dodatkowo przeprowadzić wywiad, aby oszacować procent palących z dokładnością co najwyżej $\pm\mathbf{2}\%$ przy współczynniku ufności $1-\alpha=0.95$

$n_0 = 150$ $\quad p = \frac{75}{150} = \mathbf{0.5}$ $\quad {}_{\alpha}u = 1.96$ $\quad d = 0.02$

$$n = \frac{(1.96)^2 \cdot 0.5 \cdot 0.5}{(0.02)^2} = \mathbf{2401}$$

Należy jeszcze dodatkowo rozmawiać z około $n - n_0 = 2401 - 150 = 2251$ studentami.

# Szacowanie niezbędnej liczebności próby

**Przykład c.d. (jeszcze raz to samo zadanie, ale dla innych danych):**

Przeprowadzono wywiad z 150 studentami. Z tego **75** paliło systematycznie tytoń. Z iloma studentami należy dodatkowo przeprowadzić wywiad, aby oszacować procent palących z dokładnością co najwyżej $\pm \mathbf{2.5\%}$ przy współczynniku ufności $\mathbf{1 - \alpha = 0.90}$

$n_0 = 150$ $\quad p = \frac{75}{150} = \mathbf{0.5}$ $\quad {}_{\alpha}u = 1.645$ $\quad d = 0.025$

$$n = \frac{(1.645)^2 \cdot 0.5 \cdot 0.5}{(0.025)^2} = 1082.4 \approx \mathbf{1083}$$

Należy jeszcze dodatkowo rozmawiać z około $n - n_0 = 1083 - 150 = 933$ studentami.

# Szacowanie niezbędnej liczebności próby

**Przykład c.d. (jeszcze raz to samo zadanie, ale dla innych danych):**

Przeprowadzono wywiad z 150 studentami. Z tego **105** paliło systematycznie tytoń. Z iloma studentami należy dodatkowo przeprowadzić wywiad, aby oszacować procent palących z dokładnością co najwyżej $\pm 2.5\%$ przy współczynniku ufności $1 - \alpha = 0.90$

$n_0 = 150$ $\quad p = \frac{105}{150} = 0.7$ $\quad {}_{\alpha}u = 1.645$ $\quad d = 0.025$

$$n = \frac{(1.645)^2 \cdot 0.7 \cdot 0.3}{(0.025)^2} = 876.4 \approx 877$$

Należy jeszcze dodatkowo rozmawiać z około $n - n_0 = 877 - 150 = 727$ studentami.