ROZKŁADY PRÓBKOWE

Statystyka biomedyczna

Artur Klepaczko

PARAMETRY STATYSTYKI OPISOWEJ

	Populacja	Próbka
Średnia	μ	μ_s
Proporcja	π	p
Odchylenie standardowe	$\boldsymbol{\sigma}^2$	s^2
Wariancja	$\boldsymbol{\sigma}$	s

ROZKŁAD PRÓBKOWY

- Parametry (statystyki) próbki są **zmiennymi losowymi** ponieważ zmieniają się z próbki na próbkę.

- W konsekwencji, statystyka próbki posiada swój własny rozkład, który nazywamy **rozkładem próbkowym**.

- Rozkład próbkowy, jak każdy rozkład, ma średnią i odchylenie standardowe, które w tym przypadku nazywamy **błędem standardowym**.

PRZYKŁAD

Losujemy (ze zwracaniem) dwie kule.

FIGURE 9.1 THE POOL BALLS.

TABLE 9.1 ALL POSSIBLE OUTCOMES WHEN TWO BALLS ARE SAMPLED WITH REPLACEMENT

Outcome	Ball 1	Ball 2	Mean
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

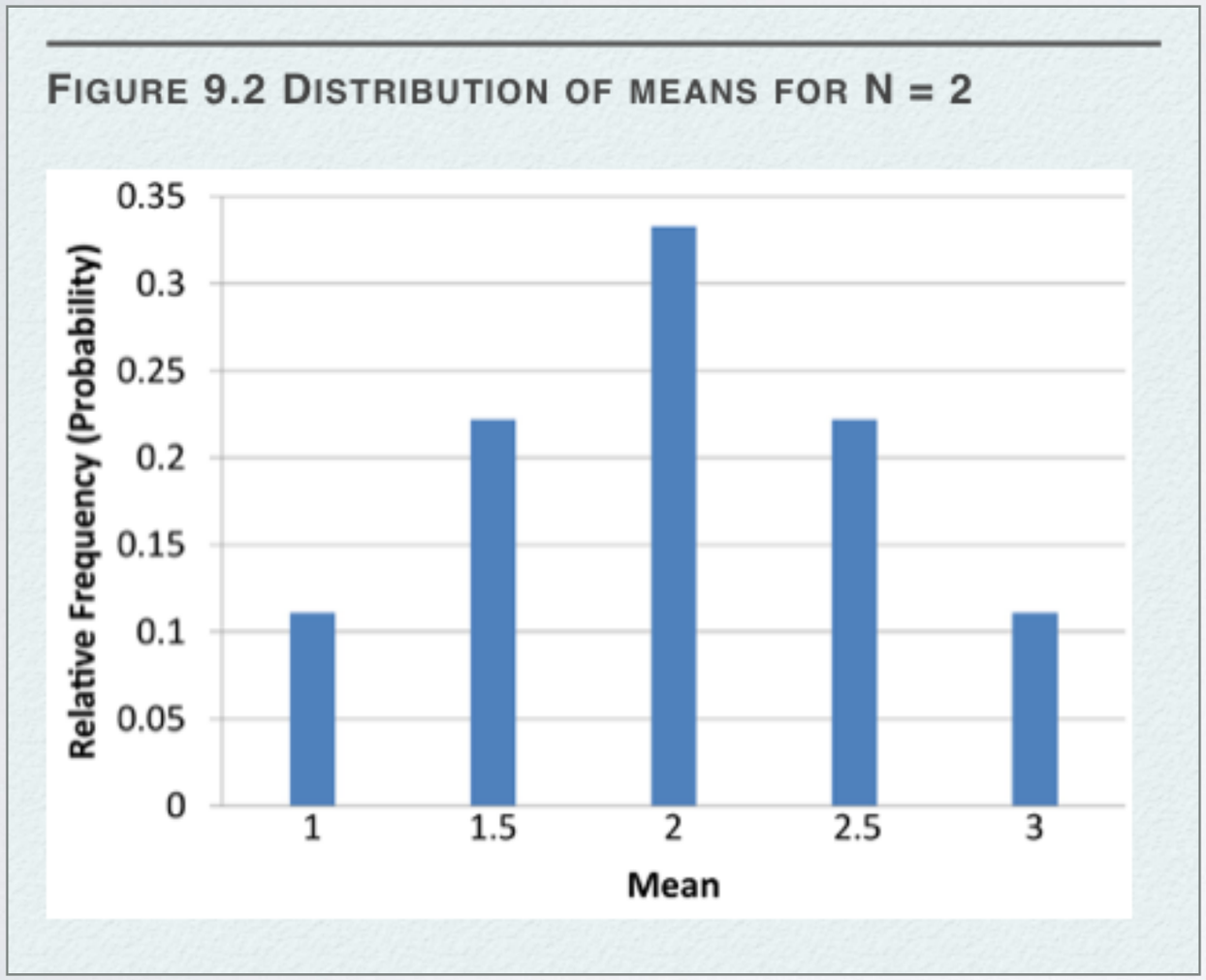
ROZKŁAD ŚREDNIEJ DLA *N*=2 KUL

TABLE 9.2 TABLE 2. FREQUENCIES OF MEANS FOR N = 2

Mean	Frequency	Relative Frequency
1.0	1	0.111
1.5	2	0.222
2.0	3	0.333
2.5	2	0.222
3.0	1	0.111

W tabeli obok przedstawiono częstotliwość wystąpienia określonej wartości średniej. Jest to również **rozkład prawdopodobieństwa** wylosowania dwóch kul, których średnia wynosi *x*. Rozkład ten nazywamy **rozkładem próbkowym średniej**.

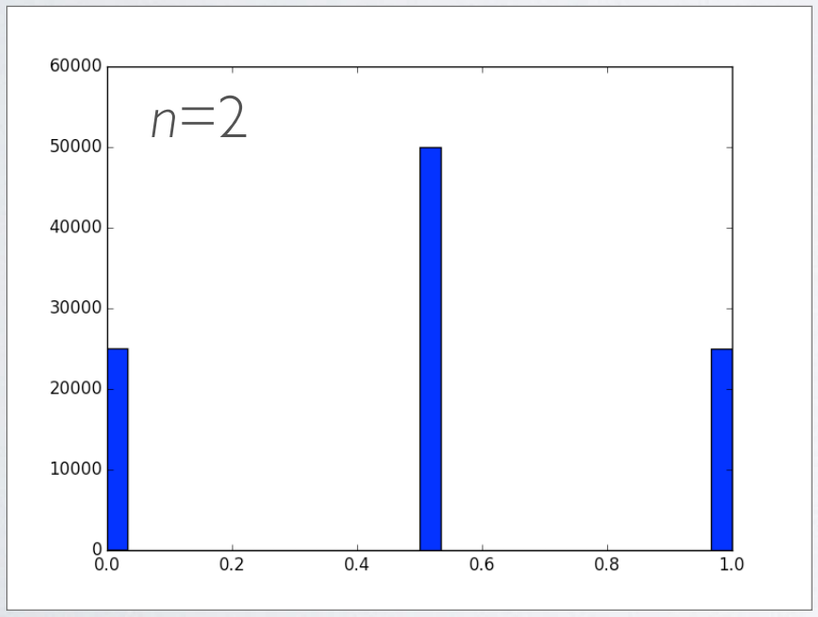
ROZKŁAD PRÓBKOWY ŚREDNIEJ

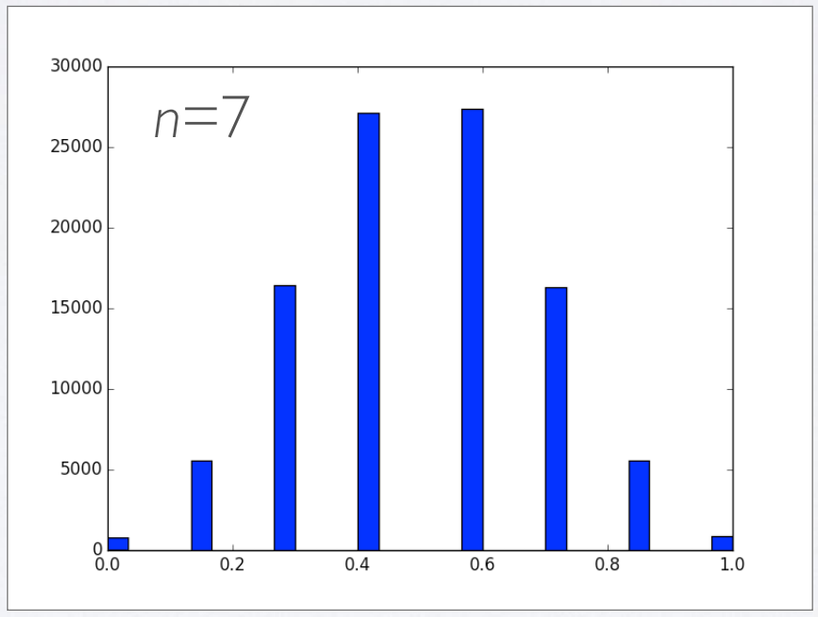


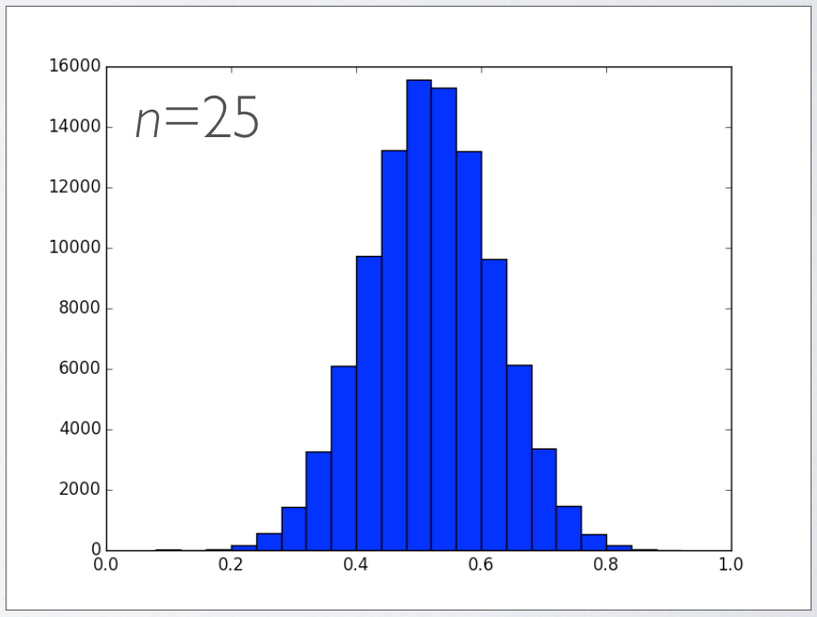
W ogólnym przypadku doświadczenie polega na losowaniu (ze zwracaniem) próbki *N* kul. Dla próbki obliczamy średnią. Losujemy drugą próbkę *N* kul i znowu liczymy i zapisujemy średnią. Powtarzamy doświadczenie *m* razy i wyznaczamy częstość (prawdopodobieństwo) wystąpienia średniej o określonej wartości. Im większe *m*, tym bardziej otrzymany rozkład będzie przypominał rozkład na rysunku.

ROZKŁAD ŚREDNIEJ

- Rzucamy monetą n razy, $n = 2, 7, 25$
- Powtarzamy eksperyment k razy, $k = 10^5$
- W każdym powtórzeniu wyznaczamy średnią liczbę wyrzuconych reszek
- Wyniki przedstawiamy na histogramie







ROZKŁAD PRÓBKOWY ZAKRESU

TABLE 9.3 ALL POSSIBLE OUTCOMES WHEN TWO BALLS ARE SAMPLED WITH REPLACEMENT

Outcome	Ball 1	Ball 2	Range
1	1	1	0
2	1	2	1
3	1	3	2
4	2	1	1
5	2	2	0
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	0

Każda statystyka próbki może mieć rozkład próbkowy…

TABLE 9.4 FREQUENCIES OF RANGES FOR N = 2

Range	Frequency	Relative Frequency
0	3	0.333
1	4	0.444
2	2	0.222

ZNACZENIE PRAKTYCZNE

- Rozkład próbkowy ma podstawowe znaczenie dla wnioskowania statystycznego.
- W przykładzie z bilami najpierw zdefiniowana była populacja, a dopiero później wyznaczaliśmy rozkład próbkowy średniej.
- W rzeczywistych badaniach proces jest odwrotny. Pozyskujemy próbkę danych, wyznaczamy dla niej średnią i zadajemy pytanie, na ile ta średnia próbkowa jest blisko średnim z innych próbek lub średniej populacji.
- Tę informację uzyskujemy bezpośrednio z rozkładu próbkowego. Przykładowo, błąd standardowy pozwala ocenić rozrzut średnich z różnych próbek. Gdy błąd ten jest mały, możemy uznać, że te średnie leżą blisko średniej populacji.

DEMO

http://onlinestatbook.com/

REGUŁA PIERWIASTKOWA

- Zależność pomiędzy zmiennością populacji a zmiennością średniej dana jest wzorem

$$SE = \frac{SD}{\sqrt{n}}$$

- Zależność ta zachodzi dla dowolnych rozkładów i ich średnich.

CENTRALNE TWIERDZENIE GRANICZNE

- Podsumowując, można zaobserwować trzy istotne właściwości rozkładu średniej:

  1) Wartość oczekiwana jest zawsze równa średniej z populacji.
  2) Błąd standardowy jest równy odchyleniu standardowemu populacji podzielonemu przez pierwiastek z *n* (liczność próby).
  3) Wraz ze wzrostem *n*, rozkład średniej zbliża się do rozkładu normalnego.

- Te trzy właściwości rozkładu średniej z próby pozostają w mocy dla **dowolnego rozkładu** (dyskretnego i ciągłego).

CENTRALNE TWIERDZENIE GRANICZNE

- Powyższy rezultat stanowi istotę tzw. *centralnego twierdzenia granicznego* (CTG), które odgrywa niezwykle ważną rolę w całej statystyce.

- W rzeczywistych problemach rzadko wiemy, jaki jest rozkład prawdopodobieństwa danych, z którymi mamy do czynienia.

- W myśl CTG, nie jest to wcale konieczne!

CENTRALNE TWIERDZENIE GRANICZNE

- Znajomość średniej oraz odchylenia standardowego rozkładu zbliżonego do normalnego pozwala wnioskować nt. zbioru danych z bardzo dużą dokładnością, a rozkład średniej z próby jest zawsze w przybliżeniu normalny

- Ale uwaga:

  - Obserwacje muszą być niezależne od siebie i reprezentatywne dla całej populacji (M. Kac)

  - CTG na zastosowanie do rozkładu średniej z próby, niekoniecznie zaś do rozkładu innych miar statystycznych

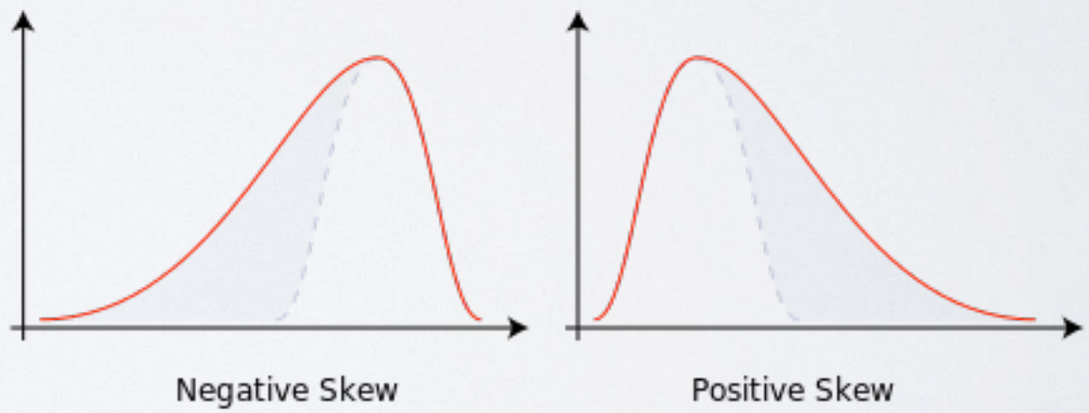
  - Minimalna liczność próby *n* potrzebna do „uzyskania" rozkładu zbliżonego do normalnego jest umowny (25, 30, 50...?).

LICZNOŚĆ PRÓBY

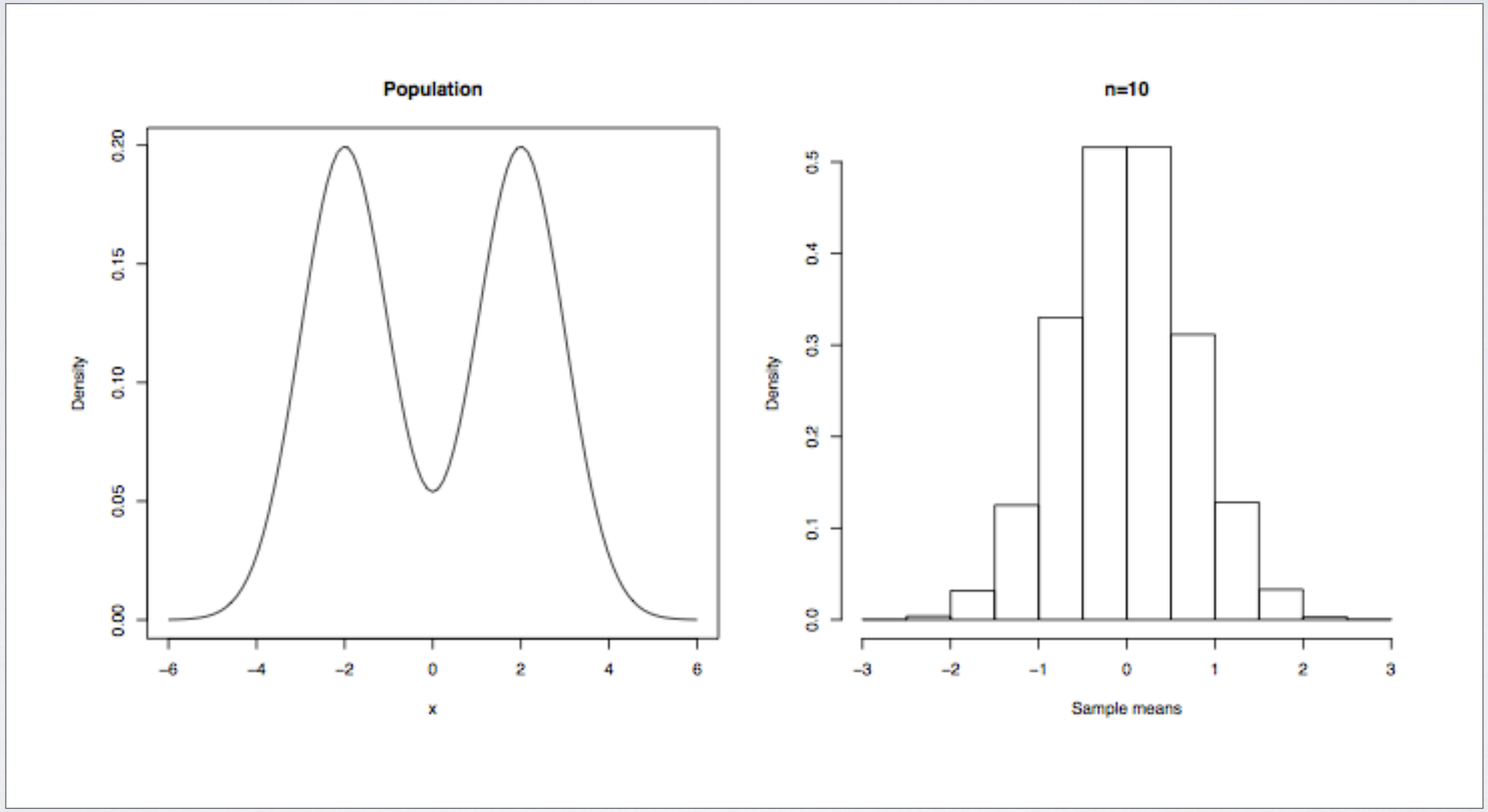
- W rzeczywistości, to czy stosujemy odpowiednio duże *n* zależy od tego, jak blisko rozkładu normalnego jest rozkład prawdopodobieństwa populacji.

- W przypadku, gdy ten rozkład jest bliski normalnemu, już *n* = 2 może być wystarczające.

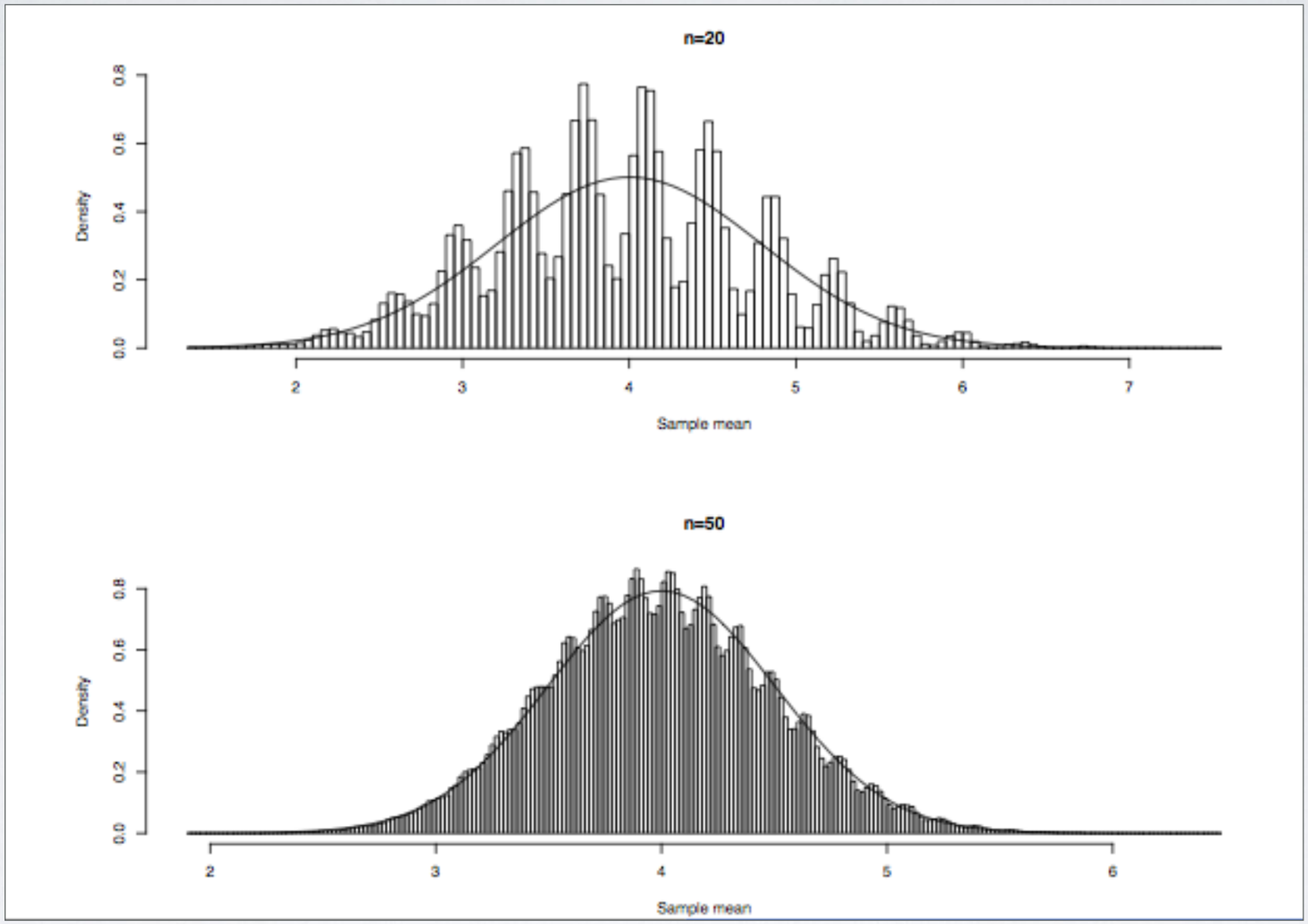
- W przypadku rozkładów o znaczącym przesunięciu wartości średniej w stosunku do wartości modalnej, n = 50 może być niewystarczające.



PRZYKŁAD #1

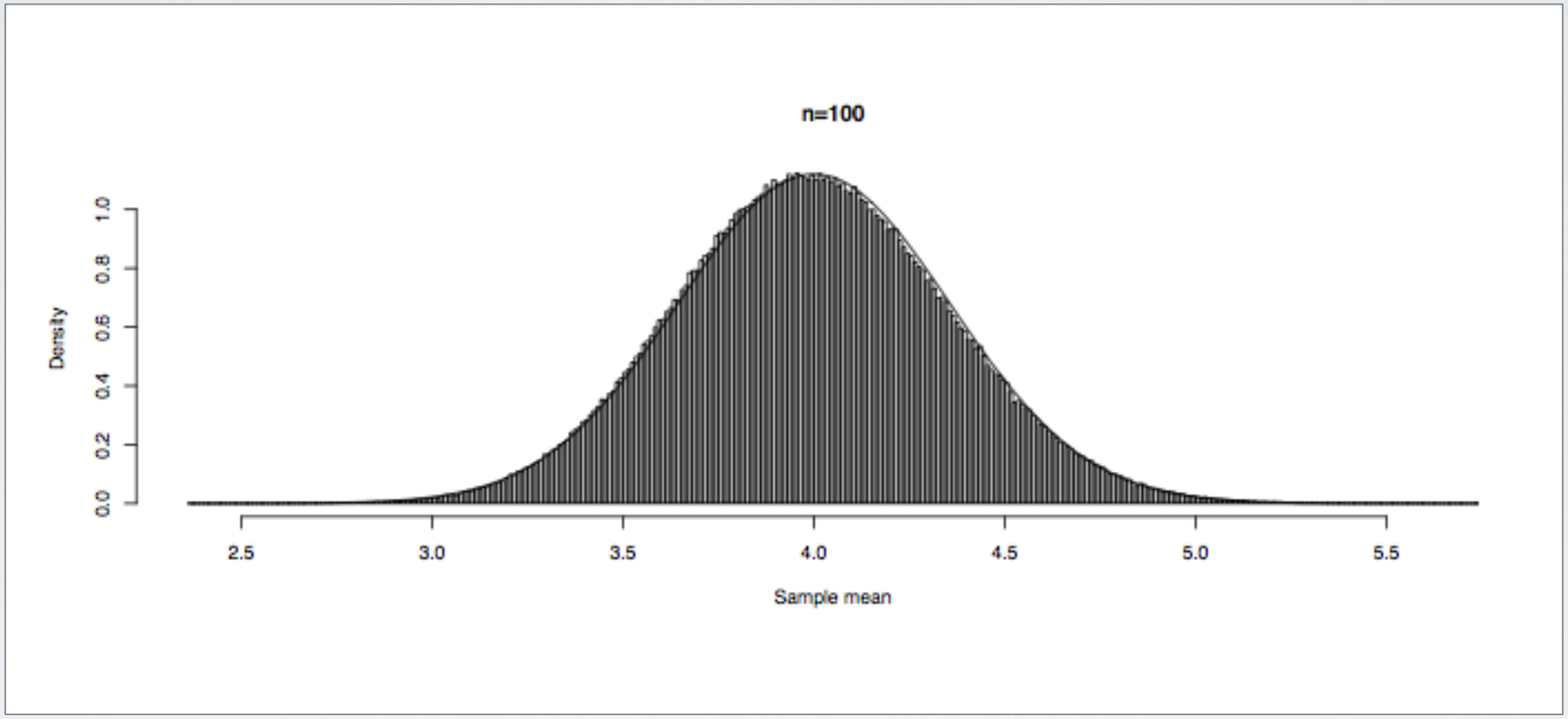


PRZYKŁAD #2



Losowanie z urny zawierającej liczby 1, 2, 9

PRZYKŁAD #2, CD.



CTG W ZASTOSOWANIACH

- Rozkład danych można wyobrazić sobie stosunkowo łatwo – obserwacje można wizualizować za pomocą wykresów, tabel, zestawień liczbowych.
- Znacznie trudniej jest mówić o rozkładzie średniej z wielokrotnej serii eksperymentów (100k rzutów monetą), ponieważ w **praktyce taki eksperyment wykonuje się raz**.
- Należy pamiętać o tym, że przy każdym takim hipotetycznym powtórzeniu, średnia (lub inny estymator) będzie inna, ponieważ inna jest próba losowa z populacji.

ROZKŁADY Z PRÓBY

- Rozkłady z próby są na ogół hipotetyczne – nie można ich wizualizować na podstawie danych pomiarowych lecz co najwyżej z zastosowaniem symulacji komputerowej.

- Statystyczne metody analizy rozkładów z próby pozwalają jednak oszacować zmienność estymatorów oraz to, czy kolejne eksperymenty potwierdzą wnioski wyciągnięte na podstawie dostępnej pojedynczej próby losowej.

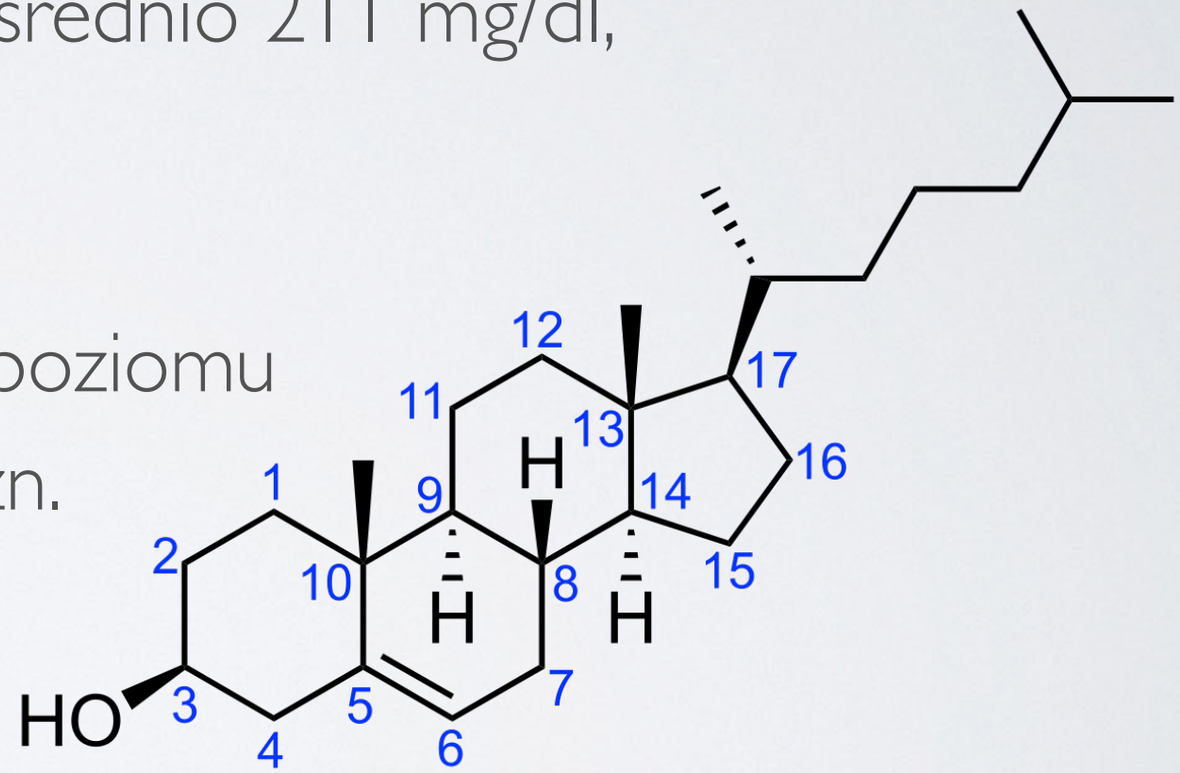
- To z kolei pozwala odpowiedzieć na pytanie, na ile dokładne jest uogólnienie wiedzy wydobytej z próby w stosunku do całej możliwej populacji.

ROZKŁAD Z PRÓBY – PRZYKŁAD #1

- Zgodnie z danymi NCHS poziom cholesterolu we krwi dla mężczyzn w USA wynosi średnio 211 mg/dl, odch. stand. 46 mg/dl.

- Eksperyment polega na pomiarze poziomu cholesterolu dla próby 25 mężczyzn.

- Jakie jest p-stwo, że średnia dla próby wyniesie powyżej 230 mg/dl?



PRZYKŁAD #1, CD.

Procedura wyznaczenia p-stwa z użyciem CTG:

1) Wyznaczyć błąd standardowy: $SE = SD/\sqrt{(n)}$

2) Wykreślić krzywą normalną aproksymującą rozkład z próby

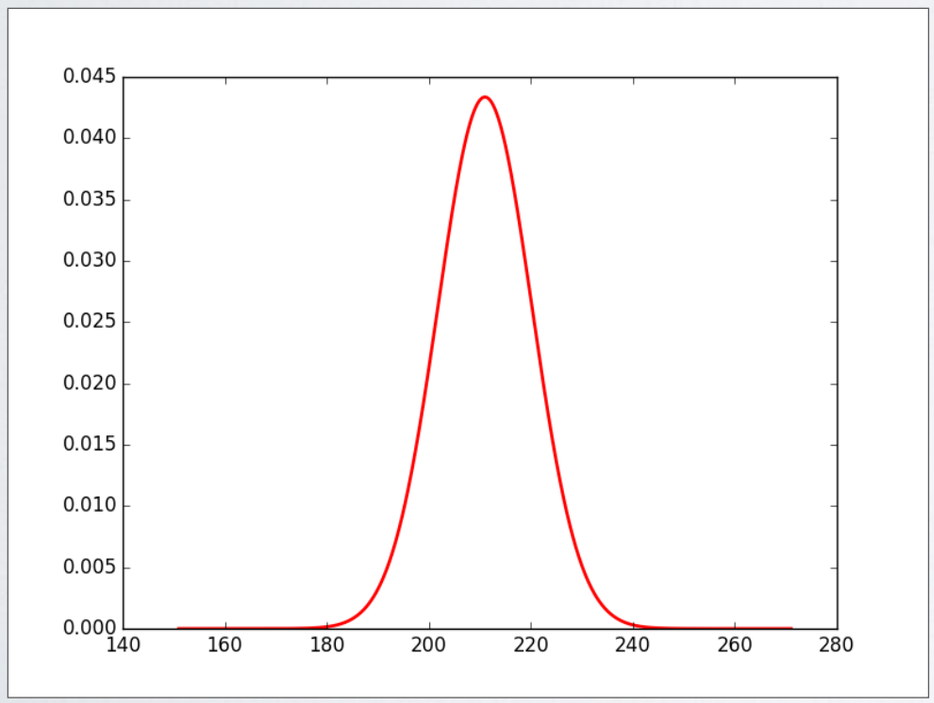
3) Ustandaryzować zmienną losową $z = (x - \mu)/SE$

4) Wyznaczyć pole powierzchni pod krzywą odpowiadające zakresowi zmiennej, dla którego chcemy wyznaczyć p-stwo (użycie tabel lub programu komputerowego)

PRZYKŁAD #1, CD.

Błąd standardowy

$$SE = \frac{SD}{\sqrt{n}} = \frac{46}{\sqrt{25}} = 9.2$$

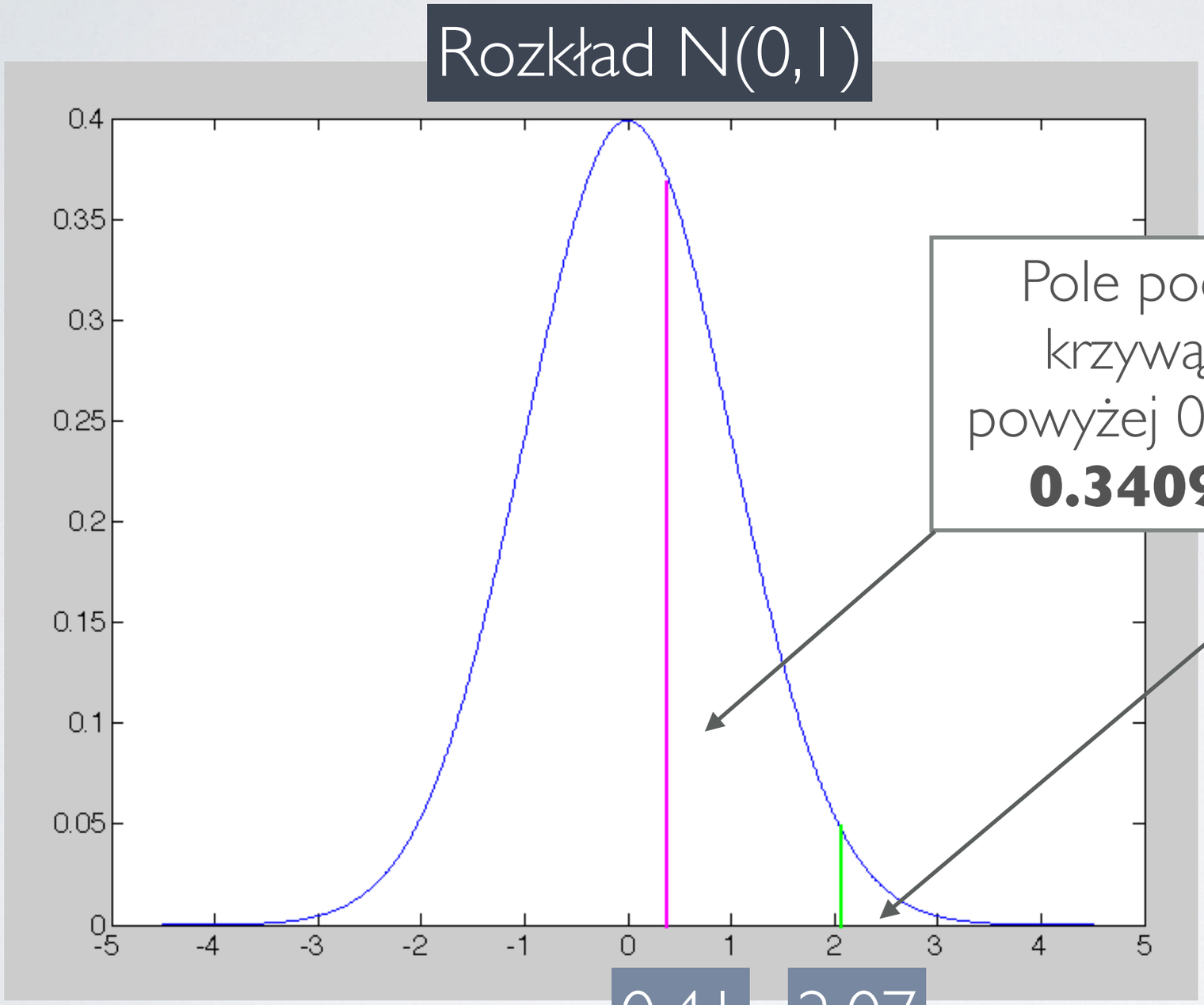


Po narysowaniu wykresu ustalamy w jakiej odległości od średniej znajduje się poszukiwana wartość poziomu cholesterolu.

$$\frac{230 - 211}{9.2} = 2.07$$

P-stwo, że w rozkładzie normalnym zmienna przyjmie wartość oddaloną o 2.07 odchylenia standardowego od średniej wynosi 1.9%.

PRZYKŁAD #1, CD.



Pole pod krzywą powyżej 0.41:
0.3409

Pole pod krzywą powyżej 2.07:
0.0192

Według rozkładu dla całej populacji:

(230 – 211) / 46
= 0.41

PORÓWNANIE Z POPULACJĄ

- Należy zauważyć, że otrzymany wynik jest znacząco różny od procentu populacji z poziomem cholesterolu powyżej 230.

- Udział ten wynosi 34% ponieważ wartość 230 jest oddalona o 0.41 odchylenia standardowego powyżej średniej.

- Średnia dla próbki jest znacznie mniej zmienna niż zmienność pojedynczych obserwacji w obrębie całej populacji.

WYZNACZANIE PRZEDZIAŁÓW UFNOŚCI

- Z zastosowaniem CTG można m.in. znacznie prościej wyznaczyć przedziały ufności estymatorów rozkładu dwumianowego dla jednorodnej próby danych nominalnych niż z użyciem wzoru Bernoulliego.

PRZYKŁAD #2 – PERCENTYLE

- Stawiamy pytanie: w jakich granicach poziomu cholesterolu znajdzie się 95% średnich z próby?

- Błąd standardowy jest równy 9.2.

- Odpytujemy pythona, ile wynoszą percentyle rzędu 2.5 oraz 97.5 dla rozkładu normalnego:

```
>>> scipy.stats.norm.ppf(0.975)
1.959963984540054
>>> scipy.stats.norm.ppf(0.025)
-1.959963845400545
```

Jak widać, dla zmiennej o rozkładzie normalnym 95% wartości mieści się w odległości 1.96 odchylenia standardowego od średniej.

PRZYKŁAD #2, CD.

- Wykonujemy obliczenia

$$211 - 1.96 \cdot 9.2 = 193.0$$

$$211 + 1.96 \cdot 9.2 = 229.0$$

- 95% średnich z próbki mieści się w przedziale od 193 mg/dl do 229 mg/dl

PRZYKŁAD #3

- Zmniejszamy liczność próby do $n = 10$ i znowu pytamy o prawdopodobieństwo tego, że średnia wartość poziomu cholesterolu w próbce będzie wynosić powyżej 230.

$n = 10$

$$SE = \frac{46}{\sqrt{10}} = 14.5$$

$n = 25$

$$SE = \frac{46}{\sqrt{25}} = 9.2$$

PRZYKŁAD #3, CD.

- Okazuje się, że teraz wartość 230 jest oddalona od średniej jedynie o 1.31 odchylenia standardowego (w stosunku do 2.07 dla $n = 25$).

- P-stwo wylosowania wartości oddalonej o 1.31 odchylenia standardowego powyżej średniej wynosi 9.6%.

- Szansa uzyskania wartości średniej z próby powyżej 230 jest więc prawie 5x większa w mniej licznej próbie (1.9% dla $n = 25$).

EKSPERYMENTY #1-3 – PODSUMOWANIE

n	SE	Przedział 95%	Szerokość przedziału
10	14.5	(182.5, 239.5)	57.0
25	9.2	(193.0, 229.0)	36.0
50	6.5	(198.2, 223.8)	25.6

- Szerokość przedziału 95% zmniejsza się – o jaki współczynnik?

PRZYKŁAD #4

- Jaka powinna być liczność próbki, aby zapewnić, że 95% średnich z próbki nie będzie odchylone od średniej z populacji bardziej niż o 5 mg/dl?

- Jak widzieliśmy, 95% obserwacji leży w zakresie ± 1.96 odchylenia standardowego od średniej, a zatem:

$$1.96 \cdot SE = 5$$

$$SE = 5/1.96$$

PRZYKŁAD #4, CD.

- Z kolei błąd standardowy jest powiązany z odchyleniem standardowym populacji regułą pierwiastkową:

$$\frac{5}{1.96} = \frac{SD}{\sqrt{n}}$$

$$\sqrt{n} = SD \cdot \frac{1.96}{5}$$

$$n = 325.1$$

- Ponieważ nie da się wylosować 325.1 osób, liczność próby musi wynieść 326.

PRZYKŁAD #5 – DLA WPRAWY

- Jaka powinna być liczność próby, aby zapewnić to, że 90% średniych z próbki nie będzie odchylona od średniej z populacji więcej niż o10 mg/dl?

- Dla rozkładu normalnego istnieje 90% p-stwo, że zmienna losowa będzie odchylona nie bardziej niż o 1.646 odchylenia standardowego od średniej.

- Czyli chcemy aby $1.645 \cdot SE = 10$

- Wykonujemy obliczenia:

$$\frac{10}{1.645} = \frac{46}{\sqrt{n}}$$

$$n = 57.3$$

Próba powinna obejmować 58 osób.