ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

23/10/2014

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Wnioskowanie statystyczne

Czyli jak bardzo jesteśmy pewni że parametr oceniony na podstawie próbki jest wartością prawdziwą dla całej populacji

Wnioskowanie statystyczne

- To jest proces formalnego wyciągania wniosków na podstawie dostępnych danych
- Na podstawie dostępnej próbki danych chcemy wyciągać wnioski dotyczące całej populacji:
  - Czy faktycznie próbka jest reprezentatywna?
  - Czy są znane/nieznane czynniki które mogą zaburzyć procedurę wnioskowania?
  - Czy jest znany bias ze względu na brakujące dane albo proces zbierania danych?
  - Czy rozumiemy „losowość” danych?
  - Czy nasze wnioskowanie dotyczy tworzenia modelu?
  - Itd..

Wnioskowanie statystyczne

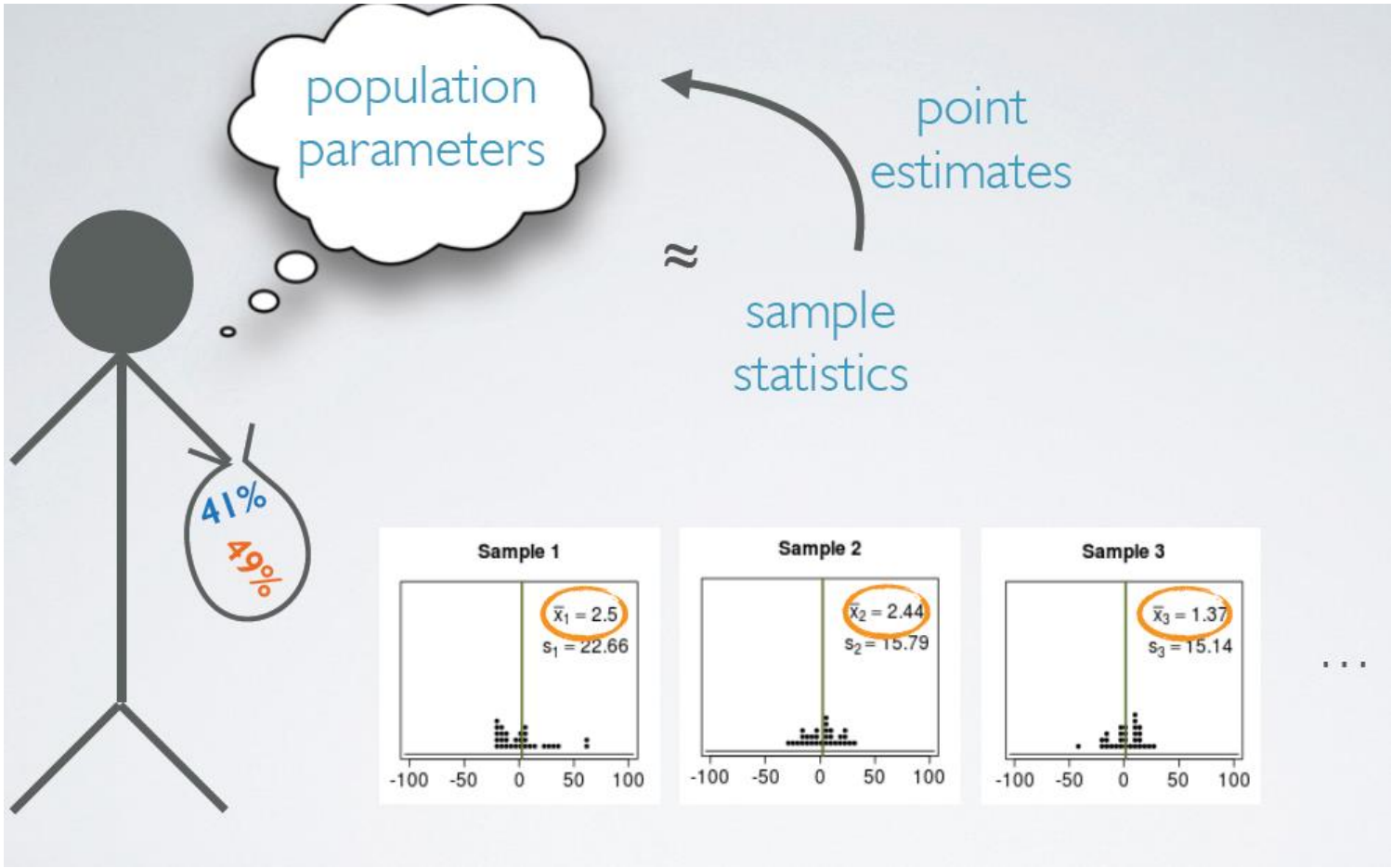
- Badania statystyczne przeprowadzone przez telefon w dniach 6-19 grudzień 2011, na próbce 2048 dorosłych (> 18 lat).
- Wyniki badań:
  - 41% sądzi że młodzi ludzie maja trudniejszą sytuację w obecnej ekonomii niż ludzie w średnim wieku lub starsi.
  - Pośród badanych w wieku 18-34 lat, 49% twierdzi że wykonują pracę której nie lubią po to aby móc się utrzymać a 12% twierdzi że pracują bez płacy aby zdobyć doświadczenie.
- Błąd statystyczny jest 2.9% dla wyników z całej próbki oraz 4.4% dla osób 18-34 lat na poziomie 95% poziomu ufności.

Wnioskowanie statystyczne

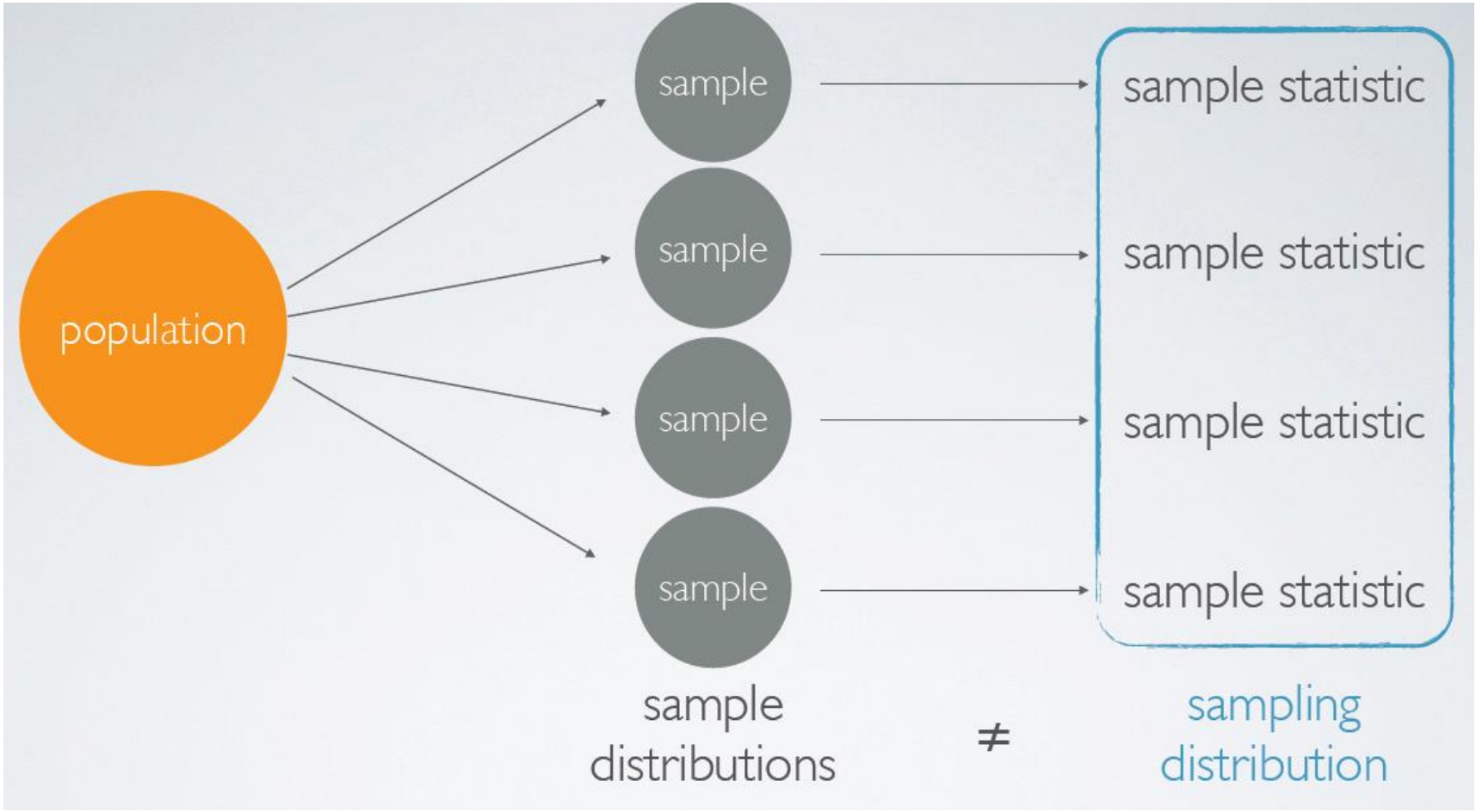
Jak to interpretujemy?

- **41% ± 2.9%:** jesteśmy przekonani na poziomie 95% przedziału ufności, że 38.1- 43.9% ludzi sądzi że młodzi ludzie maja trudniejszą sytuację w obecnej ekonomii niż ludzie w średnim wieku lub starsi.
- **49% ± 2.9%:** jesteśmy przekonani na poziomie 95% przedziału ufności, że 44.6- 53.4% młodych ludzi w wieku 18-34 lat wykonuje prace tylko po to aby pokryć koszty życia, ale jej nie lubi.

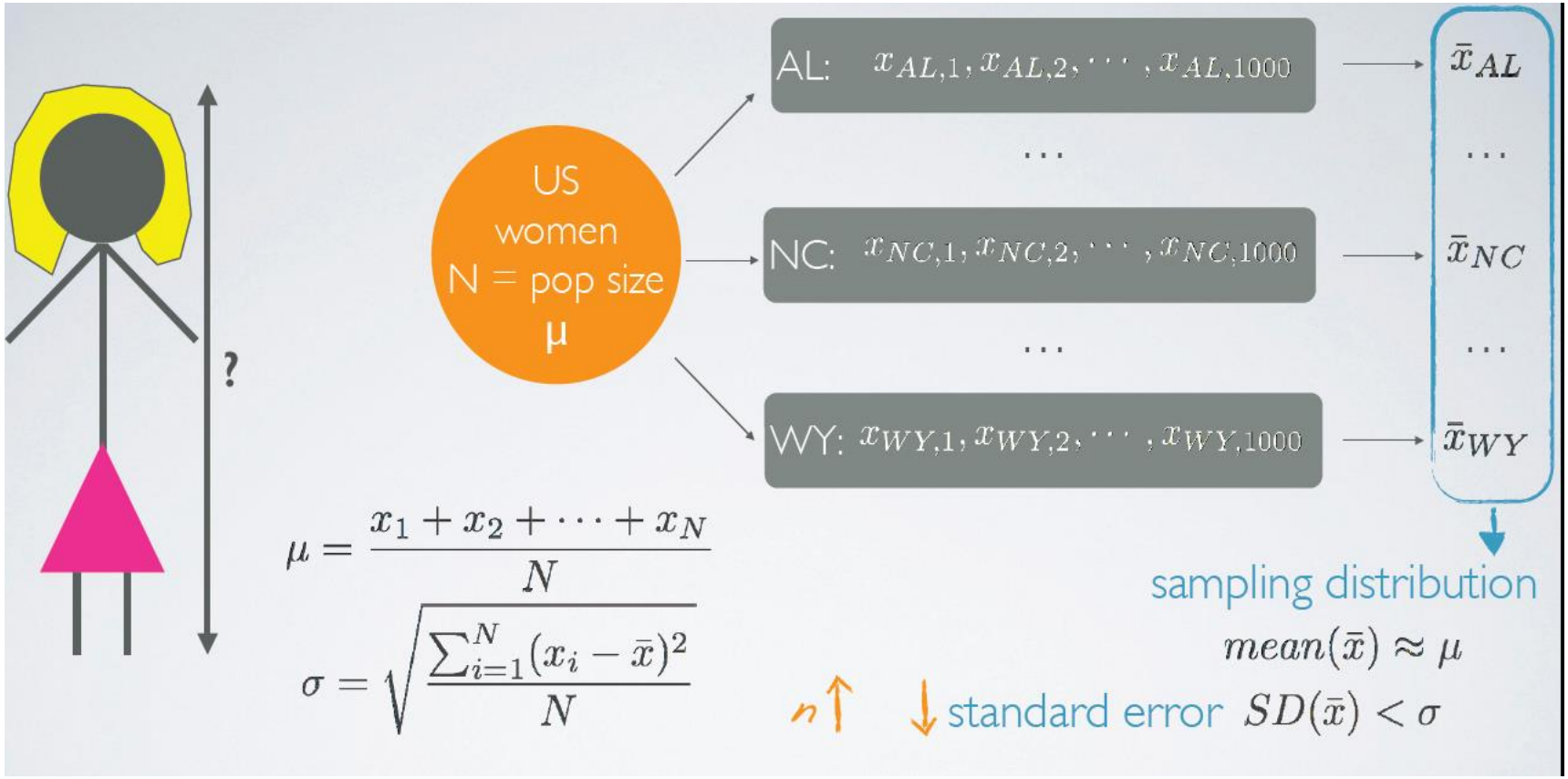
Jak przeprowadzamy takie badania?



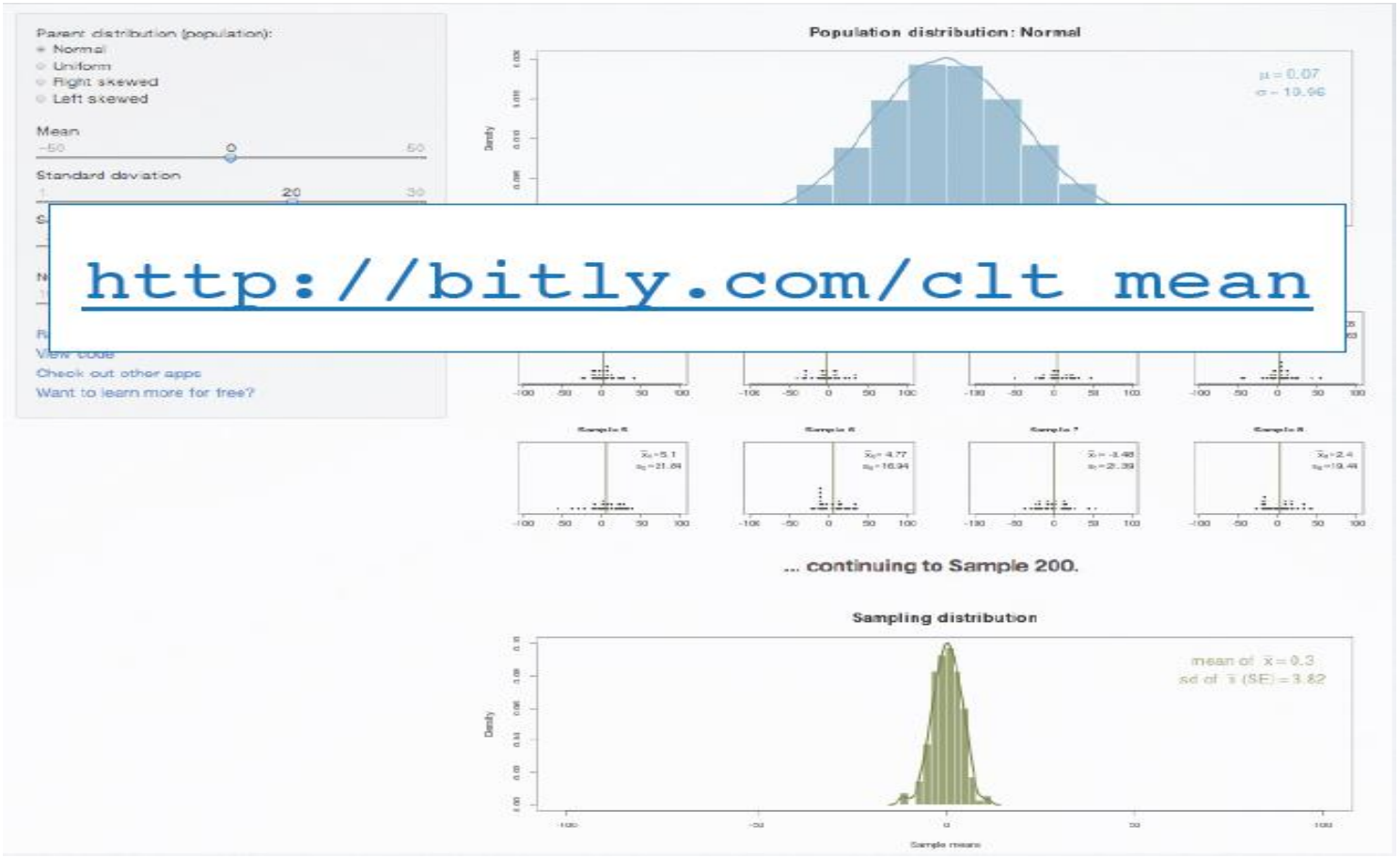
Próbka statystyczna



Przykład

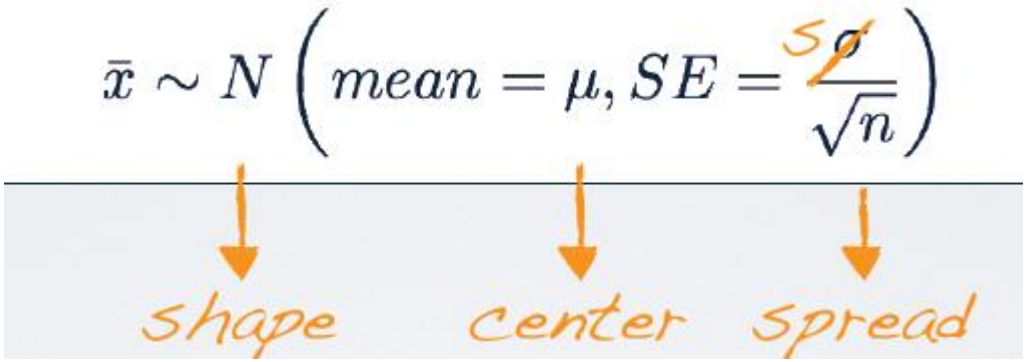

?
US
women
N = pop size
μ
AL: $x_{AL,1}, x_{AL,2}, \cdots, x_{AL,1000}$
NC: $x_{NC,1}, x_{NC,2}, \cdots, x_{NC,1000}$
WY: $x_{WY,1}, x_{WY,2}, \cdots, x_{WY,1000}$
$\bar{x}_{AL}$
$\bar{x}_{NC}$
$\bar{x}_{WY}$
$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$
$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$
sampling distribution
$mean(\bar{x}) \approx \mu$
$n \uparrow$ $\downarrow$ standard error $SD(\bar{x}) < \sigma$

Przykład


Parent distribution (population):
Normal
Uniform
Right skewed
Left skewed
Mean
Standard deviation
Check out other apps
Want to learn more for free?
Population distribution: Normal
http://bitly.com/clt_mean
... continuing to Sample 200.
Sampling distribution

Centralne twierdzenie graniczne

Rozkład zmiennej statystycznej **wartość średnia**, mierzonej na wielu próbkach tej samej populacji, ma rozkład zbliżony do normalnego

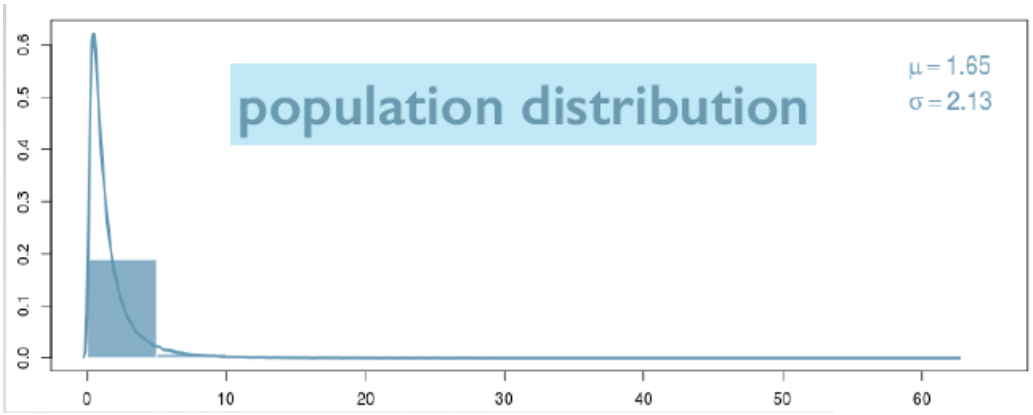


Warunki:

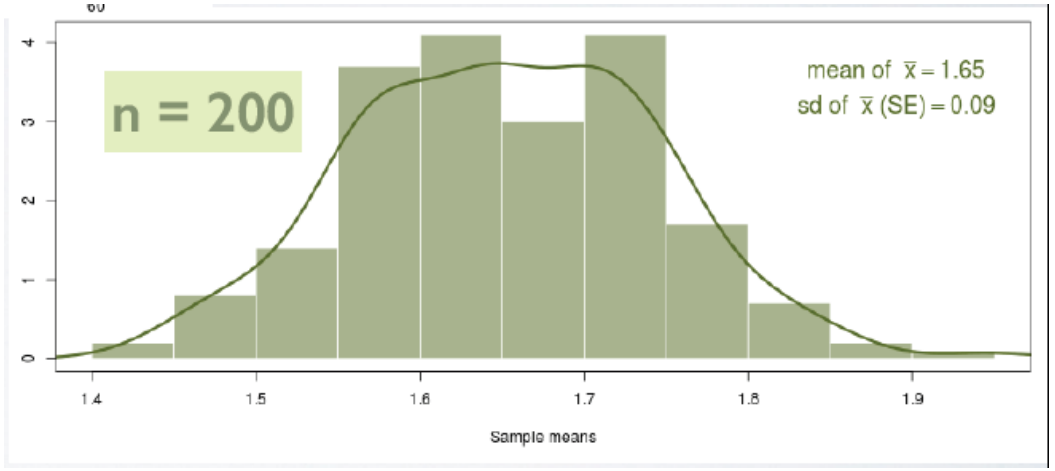
- Niezależność:
  - Próbki musza być niezależne
  - n < 10% populacji
- Skrzywienie rozkładu: albo rozkład badanej zmiennej zbliżony do normalnego, a jeżeli ma przekrzywienia to używamy duże próbki (n > 30)

Przykład:

Rozkład zmiennej w całej populacji

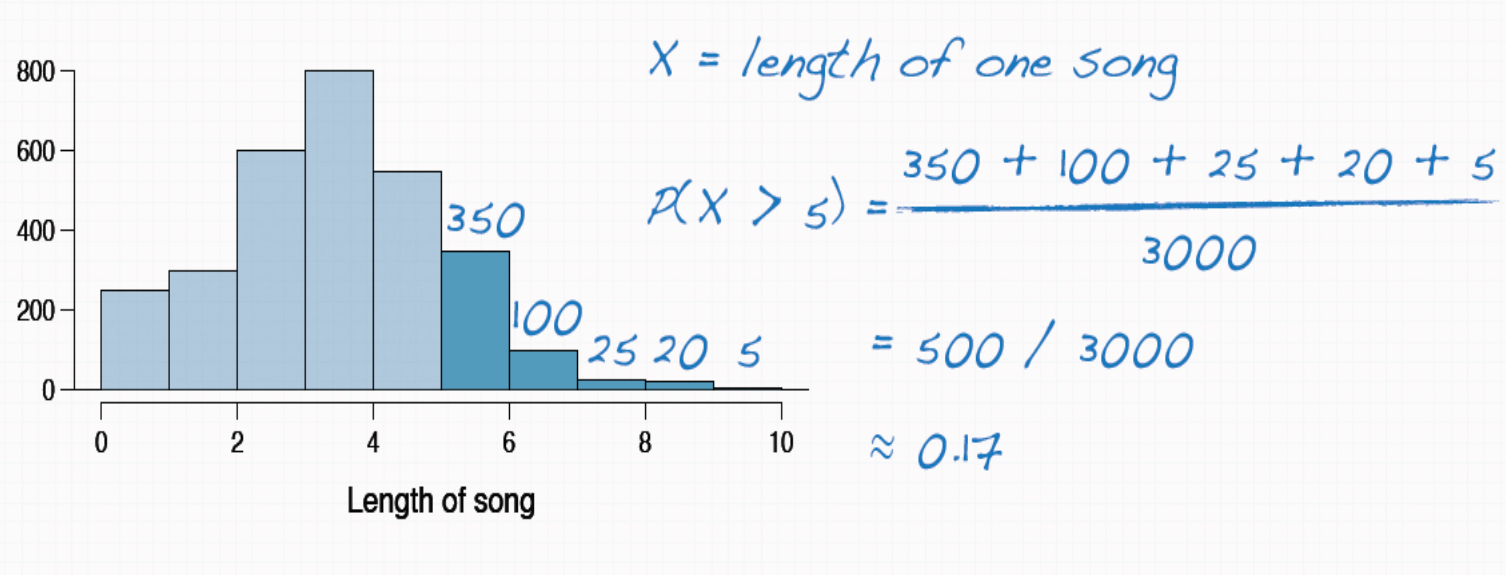


Rozkład wartości średniej dla próbek o n=200



Przykład:

Mój iPod na 3000 utworów muzycznych. Histogram pokazuje rozkład trwania tych utworów, wiem że średni czas utworu jest 3.45min, odchylenie standardowe jest 1.63min. Jakie jest prawdopodobieństwo że losowo wybrany utwór trwa dłużej niż 5min?



Przykład (cd):

Mam zamiar wybrać się w podróż która będzie trwała 6 godzin.

Wybrałam losowo 100 utworów.

Jakie jest prawdopodobieństwo że wystarczą na całą podróż?

6 hours = 360 minutes

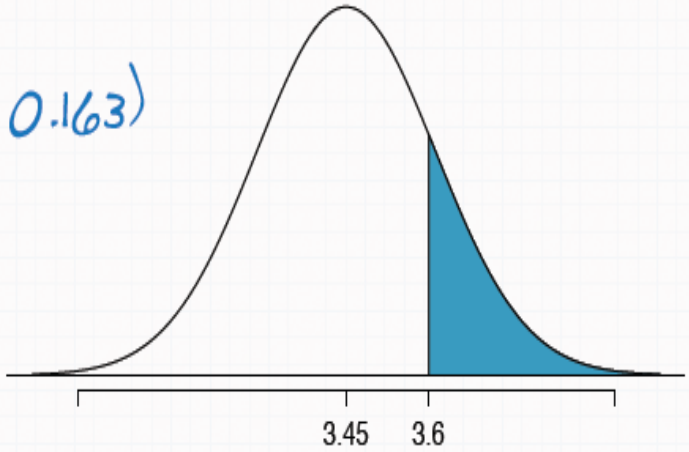
$P(X_1 + X_2 + \ldots + X_{100} > 360 \text{ min}) = ?$

$P(\bar{X} > 3.6) = ?$

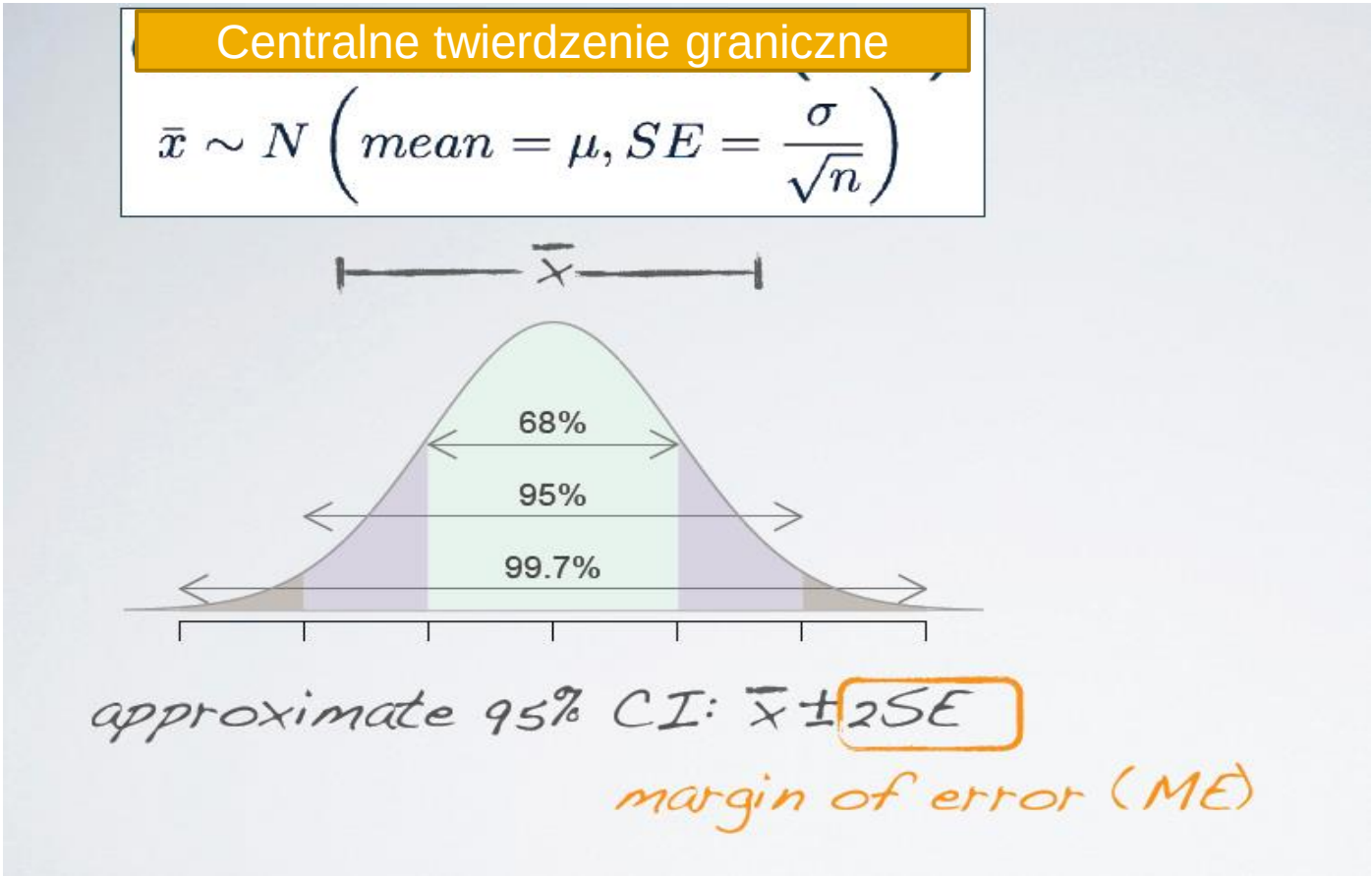
$\bar{X} \sim N(\text{mean} = \mu = 3.45, SE = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{100}} = 0.163)$

$Z = \frac{3.6 - 3.45}{0.163} = 0.92$

$P(Z > 0.92) = 0.179$



Przedział ufności (CI)



Przedział ufności (CL)

Przedział ufności dla wartości średniej

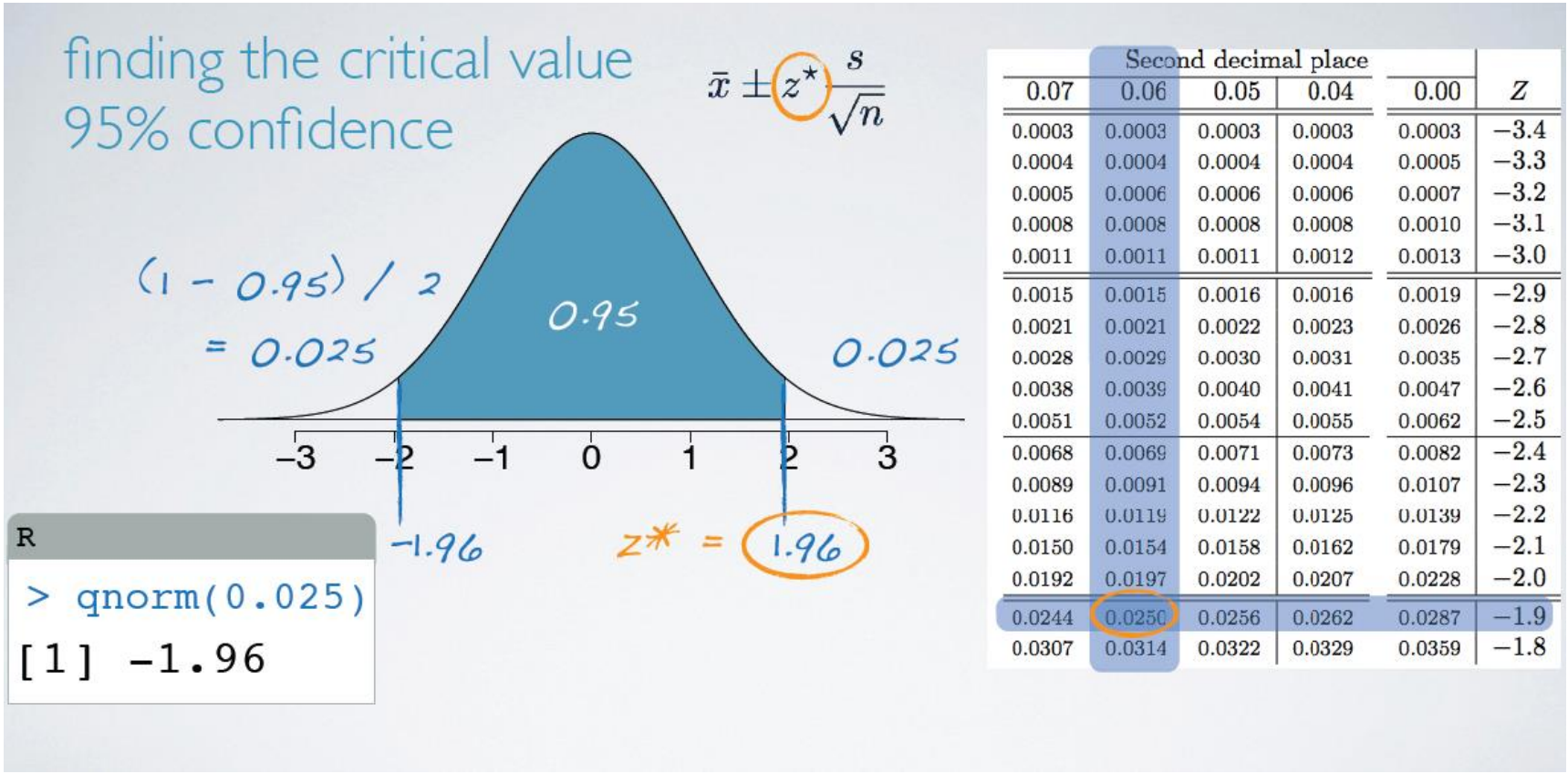
$$\bar{x} \pm z^{\star} \frac{s}{\sqrt{n}}$$

wartość krytyczna

Zawsze sprawdzamy warunki:

- Niezależność:
  - Próbki musza być niezależne
  - n < 10% populacji
- Skrzywienie rozkładu: albo rozkład badanej zmiennej zbliżony do normalnego, a jeżeli ma przekrzywienia to duże próbki (n > 30)

Wartość krytyczna



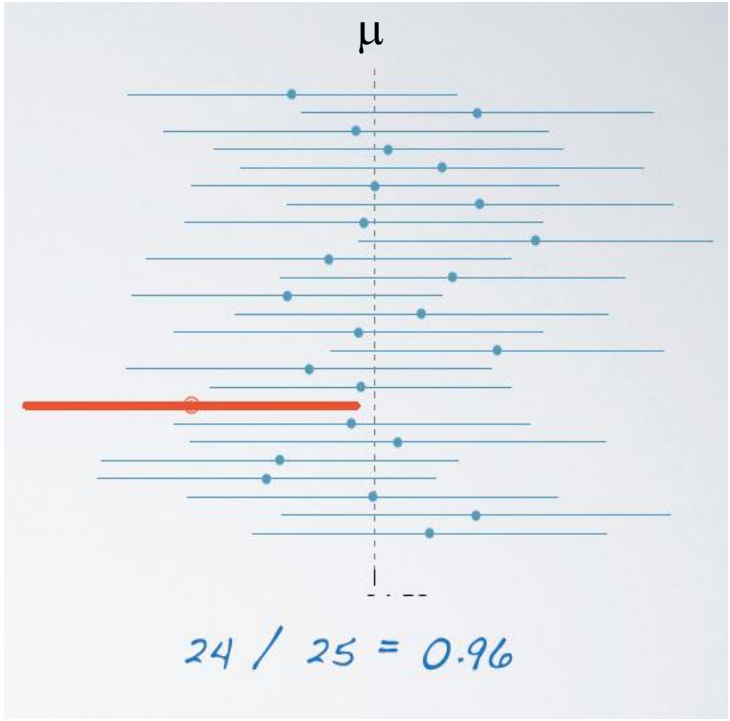
Second decimal place					
0.07	0.06	0.05	0.04	0.00	z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

Poziom ufności

Przypuśćmy że wykonaliśmy wiele pomiarów i skonstruowaliśmy dla każdego przedział ufności.

$$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

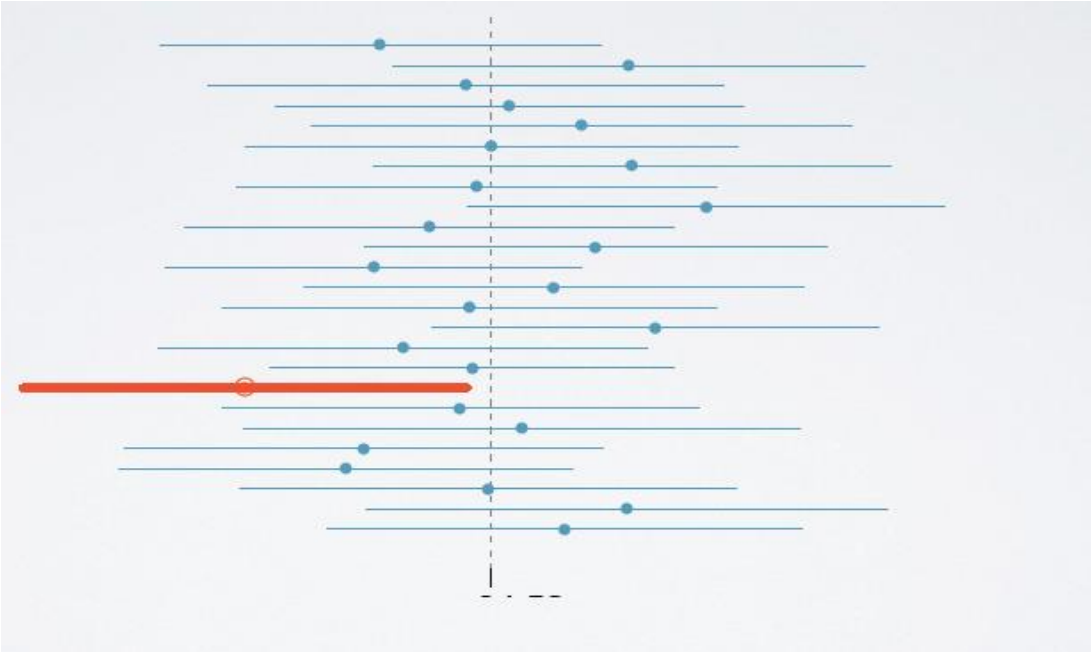
$z^* = 1.96$



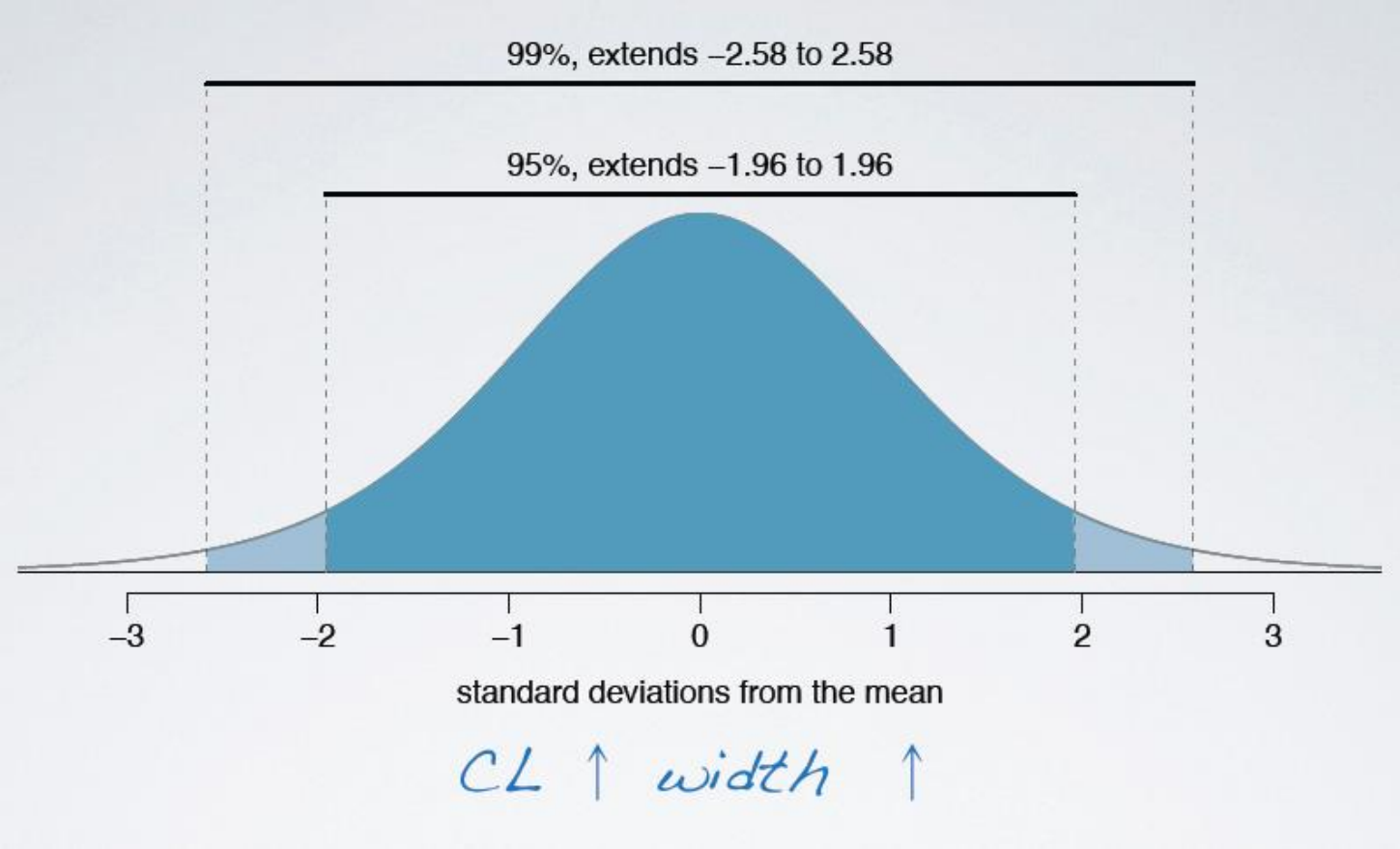
- Dla z* = 1.96 około 95% tych przedziałów będzie zawierać wartość średnią populacji (μ)
- Najczęściej używane przedziały ufności to: 90%, 95%, 98% i 99%.

Przedział ufności

Jeżeli chcemy być bardzo pewni że nasz przedział ufności będzie zawierał wartość średnia całej populacji to bierzemy szerszy przedział

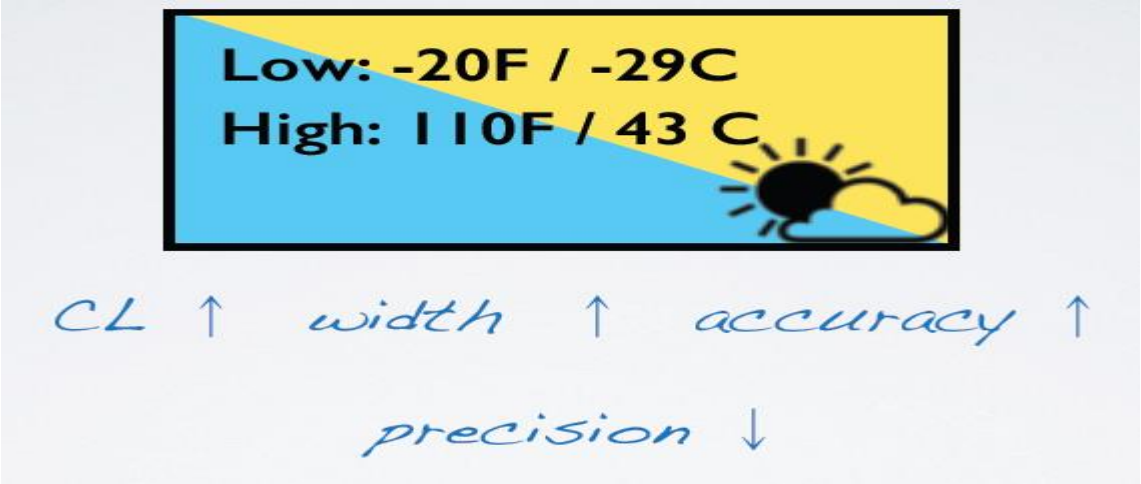


Przedział ufności



Przedział ufności

Ale nie jesteśmy wtedy precyzyjni, więc może nasza informacja jest bezużyteczna?



Precyzja vs pewność

Jak być i pewnym i precyzyjnym? Zwiększyć rozmiar próbek.

$$ME = z^{\star}\frac{s}{\sqrt{n}} \quad \rightarrow n = \left(\frac{z^{\star}s}{ME}\right)^2$$

Margin of Error

Możemy określić dopuszczalny błąd i wyznaczyc minimalną ilość elementów w próbce.

Przykład

Grupa naukowców postanowiła przebadać wpływ lekarstwa na epilepsje na inteligencję dzieci których matki brały takie lekarstwo w trakcie ciąży.

Jako miernik postanowiono przebadać IQ u 3-letnich dzieci.

Poprzednie studia wskazywały że średni IQ u 3-letnich dzieci jest równy 18 punktów. Jak dużo dzieci należy przebadać aby uzyskać 90% przedział ufności z SE ≤ 4 punkty

$ME \le 4$ pts

$CL = 90\%$

$z^* = 1.65$

$\sigma = 18$

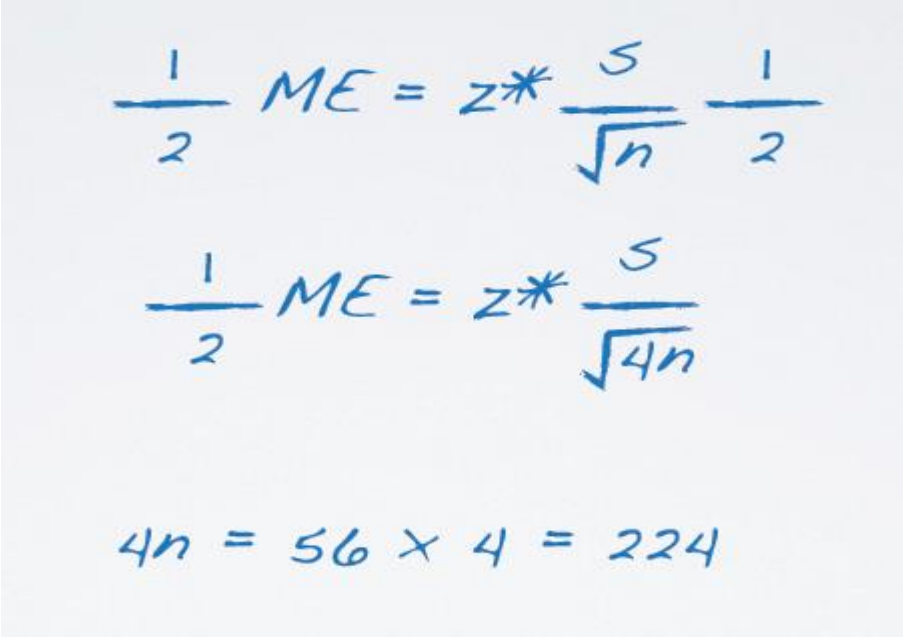
$$4 = 1.65 \frac{18}{\sqrt{n}} \rightarrow n = \left(\frac{1.65 \times 18}{4}\right)^2 = 55.13$$

Musimy przebadać co najmniej 56 dzieci

Przykład (cd)

Stwierdziliśmy że musimy przebadać co najmniej 56 dzieci.

A gdybyśmy chcieli żeby SE ≤ 2



Interpretation (raz jeszcze)

Na poziomie 95% jesteśmy pewni że zmienna x mieści się w granicach (2.72, 3.68)

$n = 50$

$\bar{x} = 3.2$

$s = 1.74$

Odchylenie standardowe

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.246$$

$$\bar{x} \pm z^* SE = 3.2 \pm 1.96\,(0.246)$$

$$= 3.2 \pm 0.48$$

$$= (2.72, 3.68)$$

Inny przykład:

$H_0: \mu = 100$

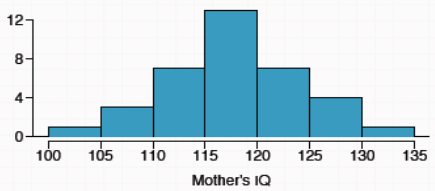
$H_A: \mu \neq 100$

$\bar{x} = 118.2$

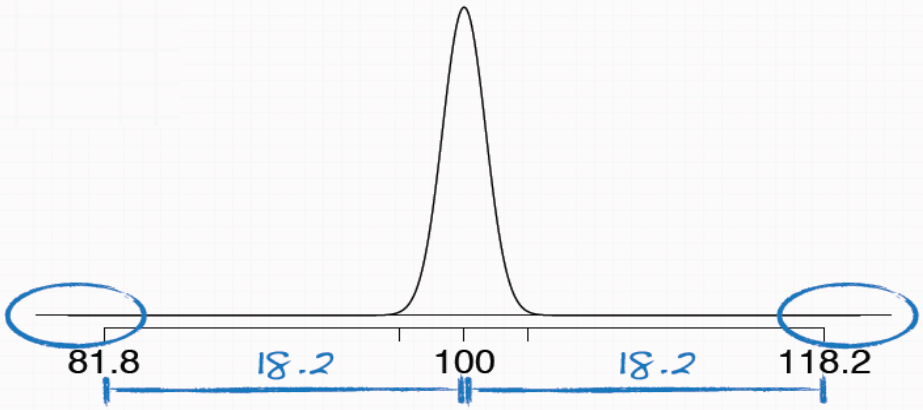
$\bar{X} \sim N(\mu = 100, SE = \frac{s}{\sqrt{n}} = \frac{6.5}{\sqrt{36}} \approx 1.083)$

$Z = \frac{118.2 - 100}{1.083} = 16.8$

$p\text{-value} \approx 0$



n	36
min	101
mean	118.2
sd	6.5
max	131



p-value bardzo małe, odrzucamy H_0

Estymatory o rozkładach prawie normalnych

Jakie estymatory mają prawie normalne rozkłady?

średnia	$\bar{x}$
różnica średnich	$\bar{x}_1 - \bar{x}_2$
proporcja w próbce	$\hat{p}$
różnica proporcji	$\hat{p}_1 - \hat{p}_2$

To są „dobre” estymatory czyli takie których rozkład ma środek w wartości tego estymatora dla całej populacji

Przedział ufności

Badanie statystyczne w 2010 roku wykazały że na 1099 przebadanych uczniów prawie 33% ogląda programy telewizyjne późno w nocy. Odchylenie standardowe tego pomiaru to 0.014.

Jaki jest 95% przedział ufności dla procentu uczniów którzy oglądają nocne programy.

$\hat{p} = 0.33$

$SE = 0.014$

$\hat{p} \pm z^* SE$

$0.33 \pm 1.96 \times 0.014$

0.33 ± 0.027

$(0.303, 0.357)$

Testowanie hipotez

Przeprowadzono badania procentowego udziału tłuszczu (BF%) na 13 601 osobach w wieku 20-80 lat. Średni BF% dla 6580 mężczyzn = 23.9 a średni BF% dla 7021 kobiet=35.0. Błąd SE różnicy między średnimi dla mężczyzn i kobiet = 0.114.

Czy dane potwierdzają tezę że średnio mężczyźni i kobiety mają różną BF%. Możesz założyć że rozkład średniego BF% jest prawie normalny.

1. Definiujemy hipotezę:

$H_0: \mu_{men} = \mu_{women}$ $H_A: \mu_{men} \neq \mu_{women}$

2. Obliczamy estymator czyli średnią różnicę:

$\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

3. Sprawdzamy warunki

Testowanie hipotez (cd)

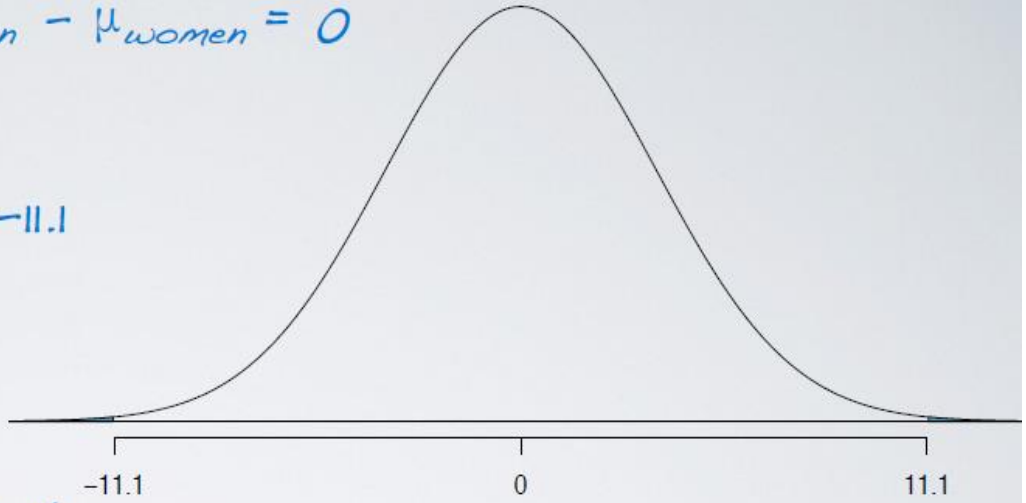
$H_0: \mu_{men} = \mu_{women} \longrightarrow \mu_{men} - \mu_{women} = 0$

$H_A: \mu_{men} \neq \mu_{women}$

$\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

$$Z = \frac{-11.1 - 0}{0.114} \approx -97.36$$

p-value $\approx 0 \longrightarrow$ Reject H_0



Te dane dają przekonywujący argument że %BF u kobiet i mężczyzn jest różny.

Podejmowanie decyzji

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✔	Type I error
	H_A true	Type 2 error	✔

- Błąd typu I : odrzucamy H_0 kiedy było prawdą
- Błąd typu II: nie odrzucamy H_0 choć prawdziwe było H_A

Podejmowanie decyzji

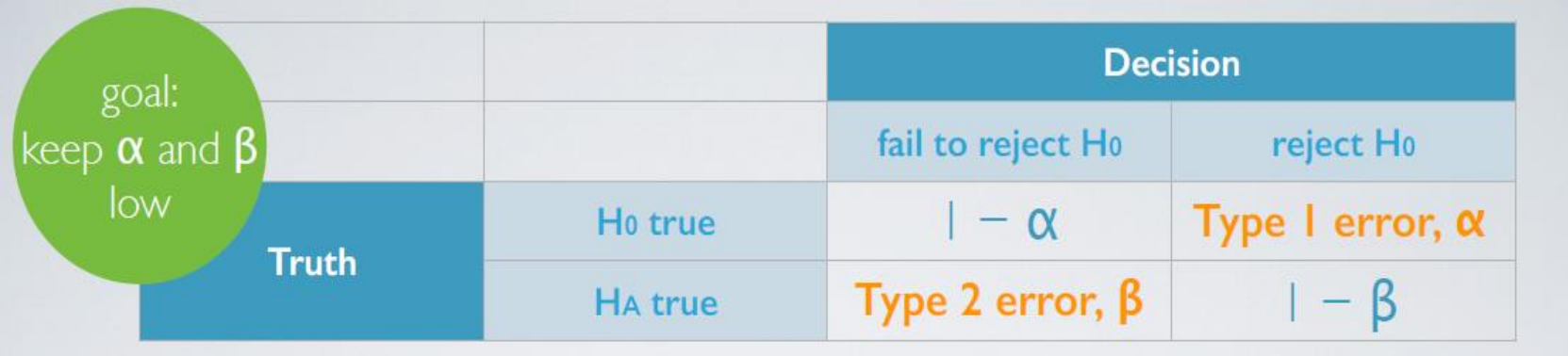
- **Błąd typu I :**
  - Odrzucamy H_0 kiedy p-value < 0.05 (α = 0.05)
  - To znaczy że jeżeli H_0 jest prawdą to nie chcemy odrzucić więcej niż 5% takich hipotez
  - Preferujemy małe wartości α, zwiększenie tej wartości powoduje zwiększenie procentu błędnych decyzji.

$$P(\text{Type I error} \mid H_0 \text{ true}) = \alpha$$

Podejmowanie decyzji

- Jeżeli błąd **typu I** jest niebezpieczny lub kosztowny wybierz małe α (np. 0.01)
- Jeżeli błąd **typu II** jest niebezpieczny lub kosztowny wybierz duże α (np. 0.10)

Podejmowanie decyzji

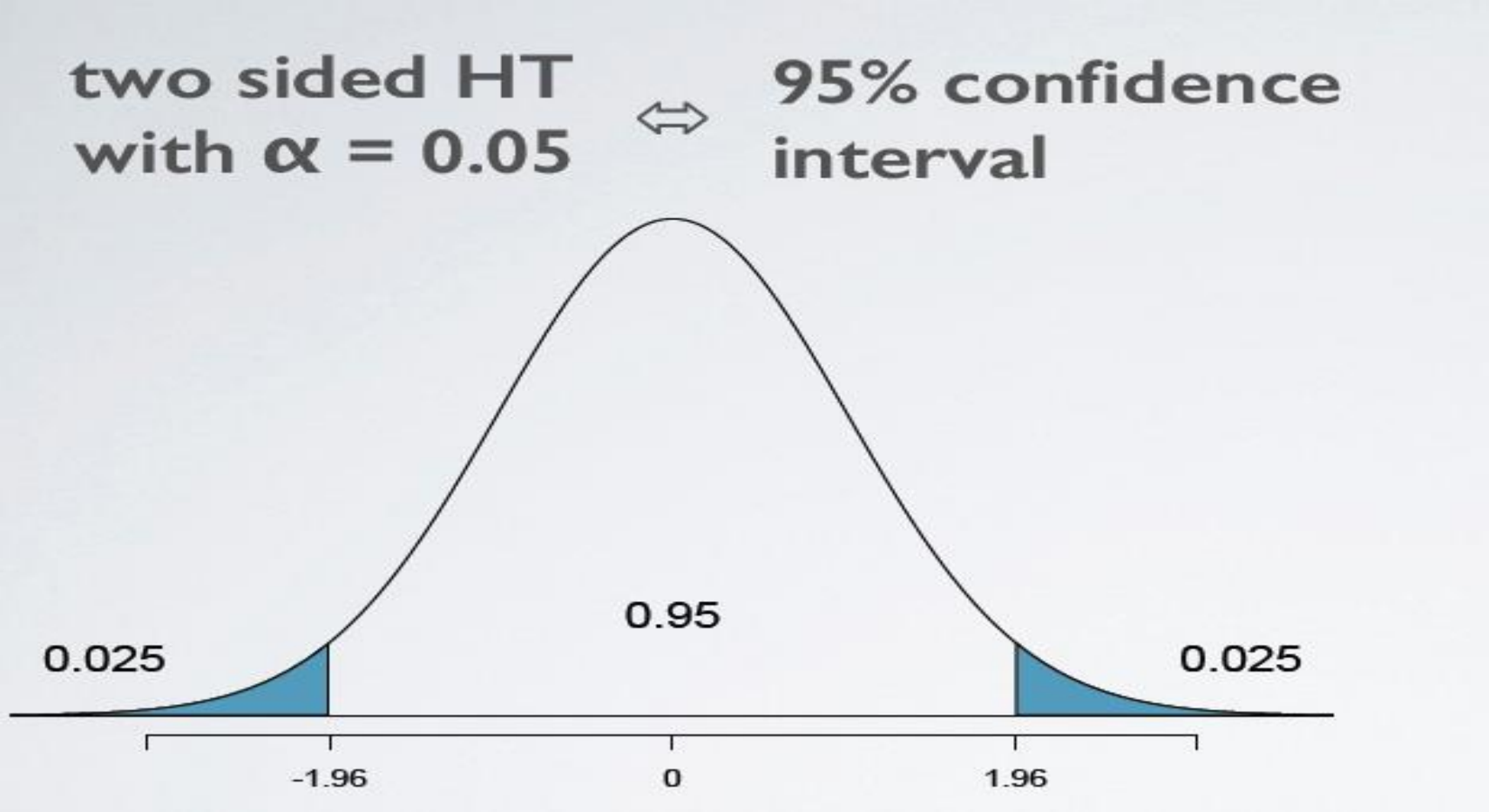


- **Błąd typu I :** odrzucamy H_0 kiedy było prawdą, prawdopodobieństwo = α
- **Błąd typu II:** nie odrzucamy H_0 choć prawdziwe było H_A, prawdopodobieństwo = β
- **Moc testu:** poprawnie odrzucać H0, prawdopodobieństwo = 1- β

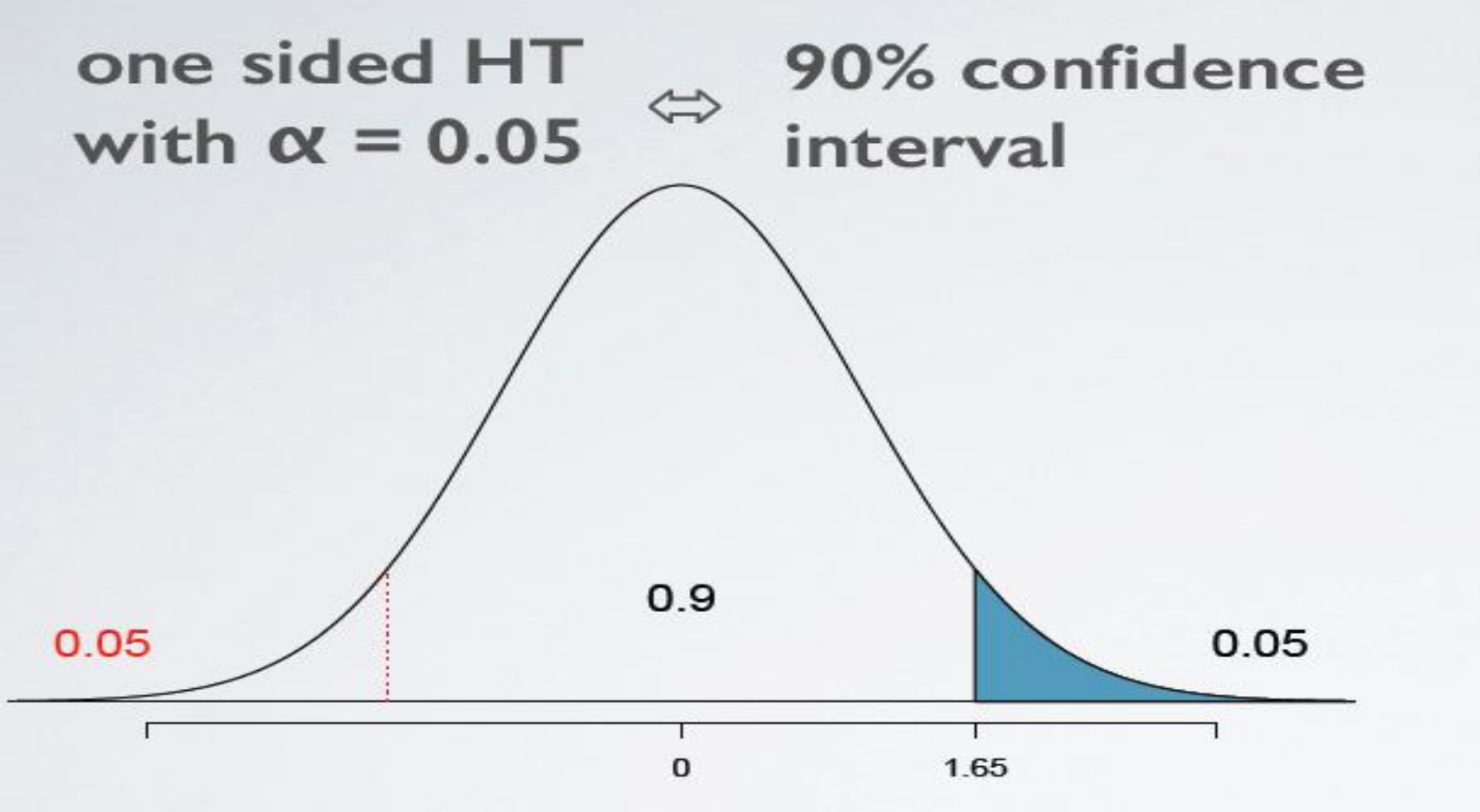
Podejmowanie decyzji

- **Błąd typu II** :
  - Prawdopodobieństwo jest trudne do policzenia
  - Jeżeli średnia populacji jest bliska wartości dla H_0 to będzie to trudne do odrzucenia
  - Prawdopodobieństwo β zależy od tej różnicy.

Przedział ufności vs znaczące statystycznie



Przedział ufności vs znaczące statystycznie



Przedział ufności i testowanie hipotezy

- Dwu-stronna hipoteza z warunkiem a jest równoważna przedziałowi ufności z CL = 1-α.
- Jedno-stronna hipoteza z warunkiem a jest równoważna przedziałowi ufności z CL = 1-(2xα)
- Jeżeli H_0 jest odrzucone to przedział ufności nie powinien zawierać wartości 0
- Jeżeli H_0 nie jest odrzucone to przedział ufnosci powinien zawierać wartość 0

Jak duże próbki ?

Dla (a) czy (b) p-wartość będzie mniejsza?

(a) n = 100
(b) n = 10,000

$$\bar{x} = 50$$
$$s = 2$$
$$H_0 : \mu = 49.5$$
$$H_A : \mu > 49.5$$

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25$$

Jak duże próbki ?

Prawdziwe różnice pomiędzy średnią z próbek i wartością dla H_0 są łatwiejsze do wykrycia dla dużych próbek.

Ale bardzo duże próbki będą wskazywać na duże znaczenie statystyczne tej różnicy, nawet jeżeli nie ma ona praktycznego znaczenia.

Czyli zaplanuj dobrze swój eksperyment

Wnioskowanie na podstawie symulacji

- Pamiętacie przykład z „promocją” kobiet i mężczyzn?

		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	total	35	13	48

% of males promoted = 21/24 ≈ 88%
% of females promoted = 14/24 ≈ 58%

Wnioskowanie na podstawie symulacji

- Wykonaliśmy symulacje przy pomocy kart. Patrząc na wyniki uznaliśmy że jest bardzo mało prawdopodobne aby przy hipotezie że nie ma „dyskryminacji” uzyskać różnicę w proporcji będącą 30% lub więcej.
- Odrzuciliśmy hipotezę H_0