ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

6/11/2014

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Wnioskowanie statystyczne dla zmiennych numerycznych

Porównywanie dwóch średnich

Boot-strapping

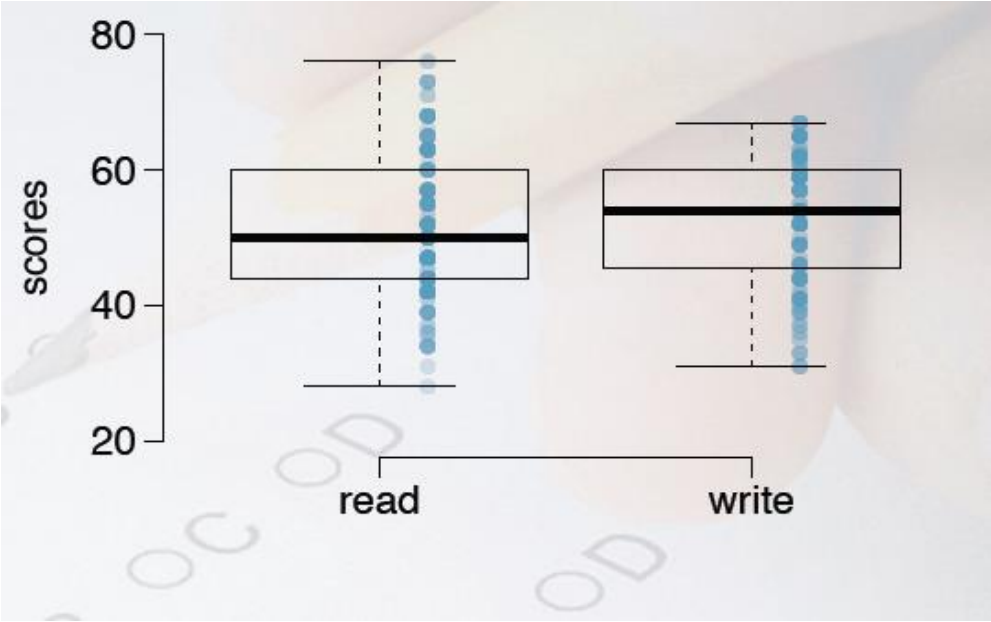
Analiza małych próbek

Analiza wariancji

Hipoteza statystyczna dla danych które występują w parach

- Wybieramy losowo 200 studentów którzy piszą testy pisemne oraz zdaja egzamin ustny.
- Czy porównywalne są wyniki uzyskane w egzaminach pisemnych i ustnych?

	ID	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
...	...	...	...
200	137	63	65



Analizujemy danych w „parach”

- Mamy dwa zbiory danych o których wiemy że nawzajem tworzą pary: każdy student pisał egzamin pisemny i odpowiadał ustnie
- W takiej sytuacji wygodnie jest monitorować średnia różnicę

  diff = read – write

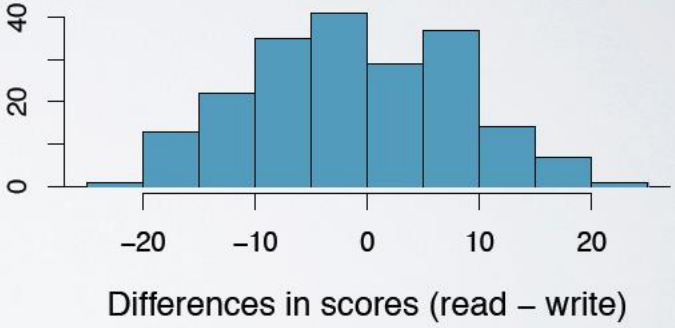
	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
…	…	…	…	…
200	137	63	65	-2

Analizujemy danych w „parach”

- Chcemy ocenić: średnia różnica dla całej populacji studentów: μ_{diff}
- Mamy dostępne wyniki badań dla próbki losowo wybranych 200 studentów: $\bar{x}_{diff}$

	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
…	…	…	…	…
200	137	63	65	-2

$$\bar{x}_{diff} = -0.545$$
$$s_{diff} = 8.887$$
$$n_{diff} = 200$$



Analizujemy danych w „parach”

- Formułujemy hipotezę: efektywnie mamy tylko jedna zmienną numeryczną

diff
5
11
19
-5
…
-2

$$H_0 : \mu_{diff} = 0$$

Nie ma różnicy pomiędzy średnią oceną za egzamin pisemny i ustny

$$H_A : \mu_{diff} \neq 0$$

Jest różnica pomiędzy średnią oceną za egzamin pisemny i ustny

Analizujemy danych w „parach”

- Przyjmujemy założenie o rozkładzie zbliżonym do normalnego.

$H_0 : \mu_{diff} = 0$

$H_A : \mu_{diff} \neq 0$

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

$n_{diff} = 200$

$\bar{X}_{diff} \sim N(mean = 0, SE = \frac{8.887}{\sqrt{200}} \approx 0.628)$

Testowanie hipotezy statystycznej

- Zdefiniowanie hipotezy: $H_0 : \mu_{diff} = 0$, $H_A : \mu_{diff} \neq 0$
- Obliczenie estymator punktowy: $\overline{x}_{diff}$
- Sprawdzenie warunków:
  - Niezależność: n < 10%
  - Rozmiar próbki i skrzywienie: n > 30
- Narysowanie rozkładu zmiennej, zaznaczenie obszaru p-value, obliczenie wartości testu statystycznego
- Podjęcie decyzji w sprawie przyjęcia lub odrzucenia hipotezy H_o

Analiza danych występujących w „parach”

$H_0 : \mu_{diff} = 0$

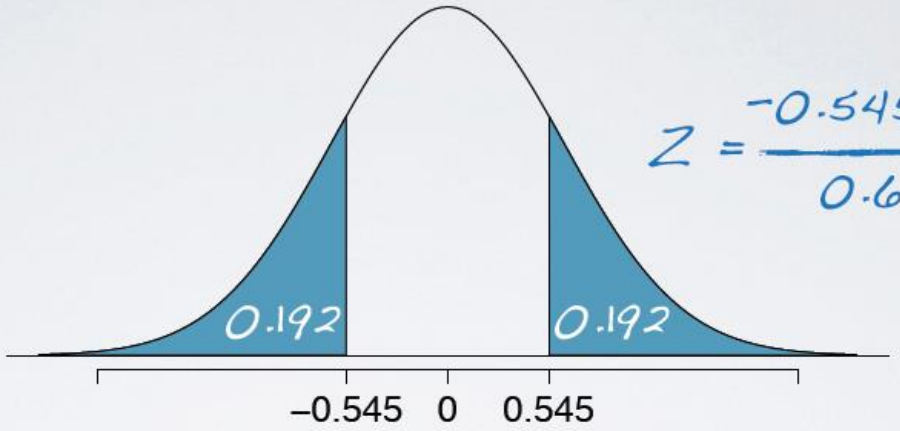
$H_A : \mu_{diff} \neq 0$

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

$n_{diff} = 200$

$\bar{x}_{diff} \sim N(mean = 0, SE = 0.628)$



$$Z = \frac{-0.545 - 0}{0.628} = -0.87$$

p-value = 0.192 × 2

= 0.384

https://bitly.com/dist_calc

p-value: prawdopodobieństwo aby na przebadanych losowo 200 studentów otrzymana średnia wyników egzaminu pisemnego i ustnego była 0.545 w sytuacji kiedy faktyczna średnia różnica jest równa zero.

Analiza danych występujących w „parach”

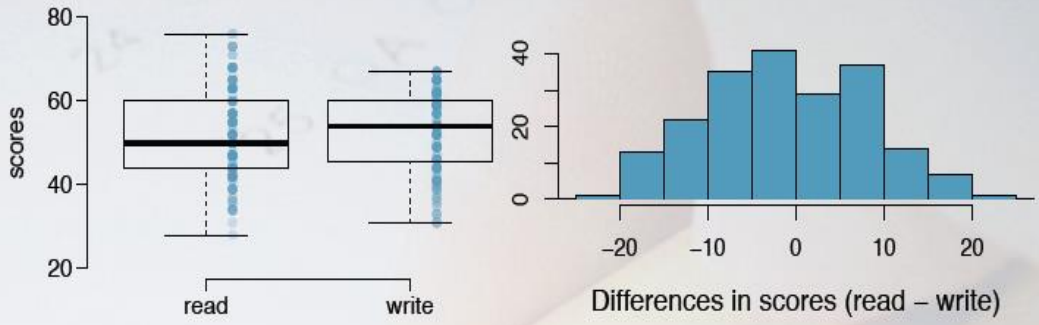
	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
…	…	…	…	…
200	137	63	65	-2

dependent

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

$n_{diff} = 200$



$H_0 : \mu_{diff} = 0$

$H_A : \mu_{diff} \neq 0$

$p - value = 0.384$

Fail to reject H_0

Testowanie przedziału ufności

$$\bar{x}_{diff} \pm z^{\star} SE_{\bar{x}_{diff}}$$

$$\bar{x}_{diff} \pm z^{\star} \frac{s_{diff}}{\sqrt{n_{diff}}}$$

Testowanie przedziału ufności

- Czy oczekujemy że 95% przedział ufności średniej różnicy ocen będzie zawierał wartość 0 ? TAK!
- Wyznacz 95% przedział ufności

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

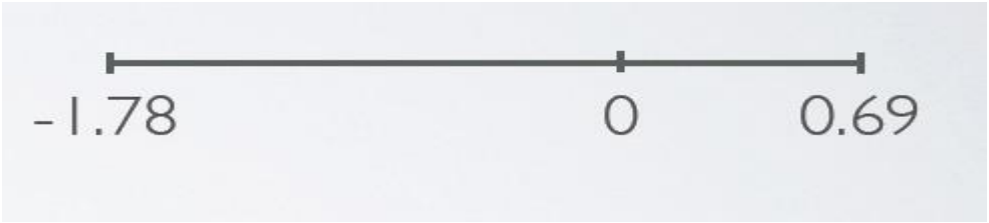
$n_{diff} = 200$

$SE = 0.628$

$\bar{x} \pm z^* SE = -0.545 \pm 1.96 \times 0.628$

$= -0.545 \pm 1.23$

$= (-1.78, 0.69)$



Analizujemy niezależne średnie

point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm z^* SE_{\bar{x}_1 - \bar{x}_2}$$

Standard error of difference between two independent means:

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Sprawdzamy warunki

- Warunki dla porównywania dwóch niezależnych średnich
  - Niezależność
    - Wewnątrz jednej grupy
      - Próbka losowa
      - Nie więcej niż 10% całej populacji
    - Pomiędzy grupami
      - Grupy musza być niezależne, nie być parą
  - Rozmiar/skrzywienie próbki
    - Jeżeli duże skrzywienie rozkładu każda próbka powinna mieć n > 30

Testowanie różnicy między niezależnymi średnimi

- Hipoteza zerowa: $H_0 : \mu_1 - \mu_2 = 0$
- Alternatywna hipoteza: $H_A : \mu_1 - \mu_2 \neq 0$

- Kolejne kroki tak jak poprzednio

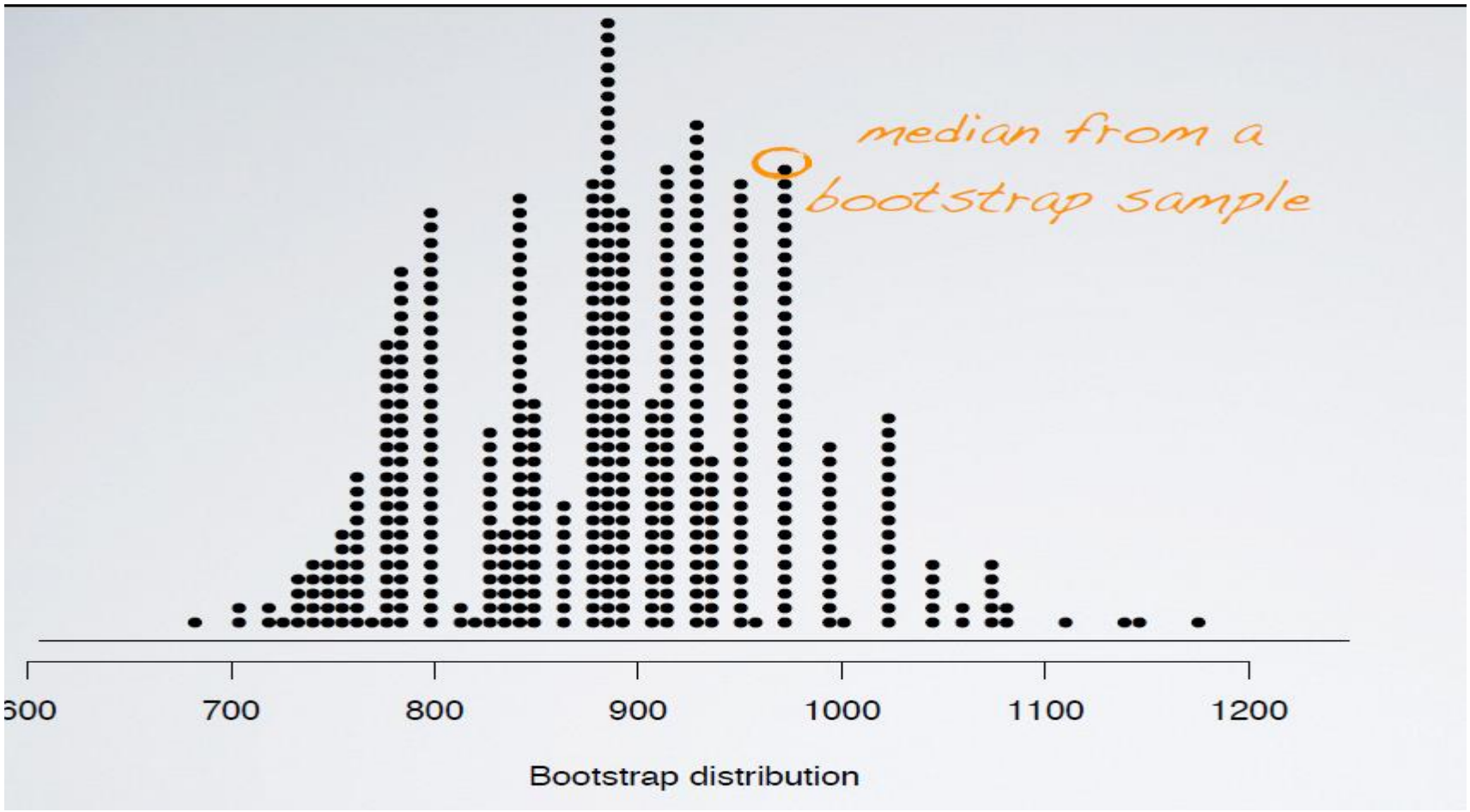
Bootstrapping

- Mamy do dyspozycji tylko jedną próbkę

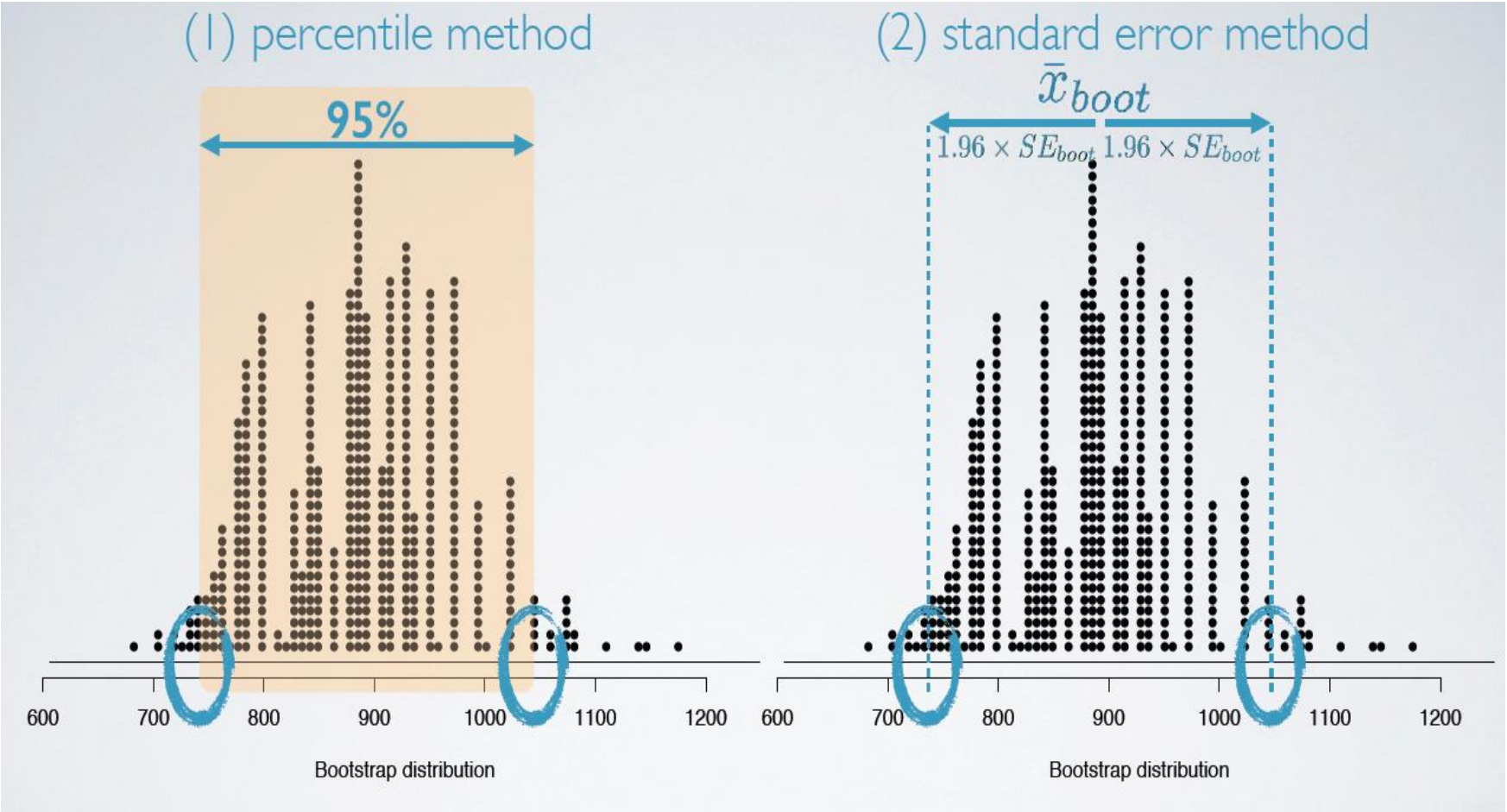
1) Losujemy ze zwracaniem z tej próbki i otrzymujemy próbkę o tej samej ilości elementów co pierwotna
2) Wyliczamy statystykę dla nowej próbki: średnia, mediana, proporcja

- Powtarzamy (1)+(2) i badamy rozkład statystyki próbek bootstrap czyli np. średniej

Bootstrapping

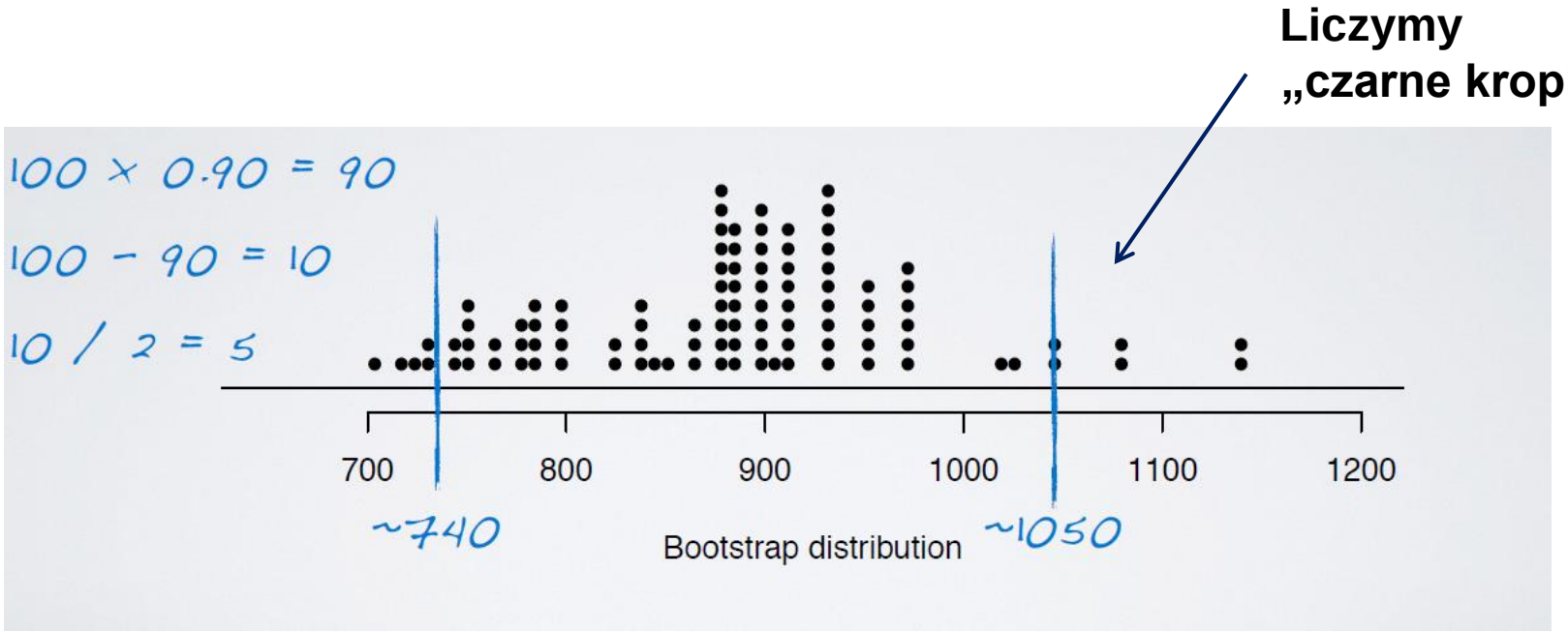


Bootstrapping



Bootstrapping

- 90% boostrap przedział ufności dla 100 bootstrap próbek: percentile metoda



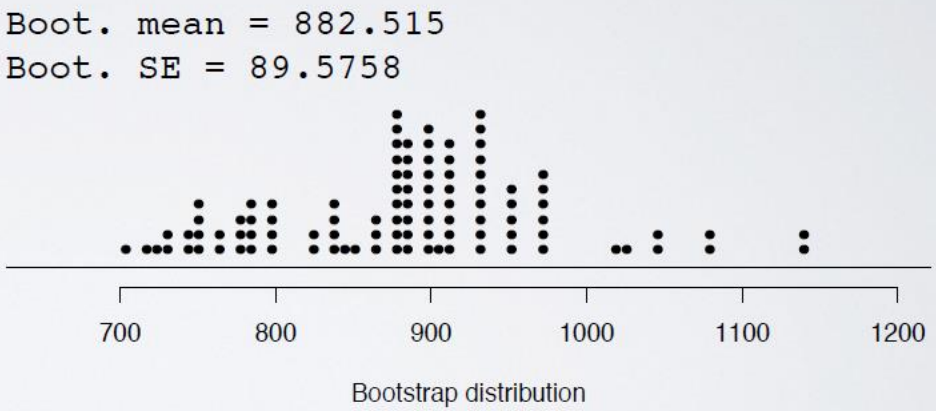
Bootstrapping

- 90% boostrap przedział ufności dla 100 bootstrap próbek: SE metoda

$\bar{x}_{boot} \pm z^* SE_{boot} =$

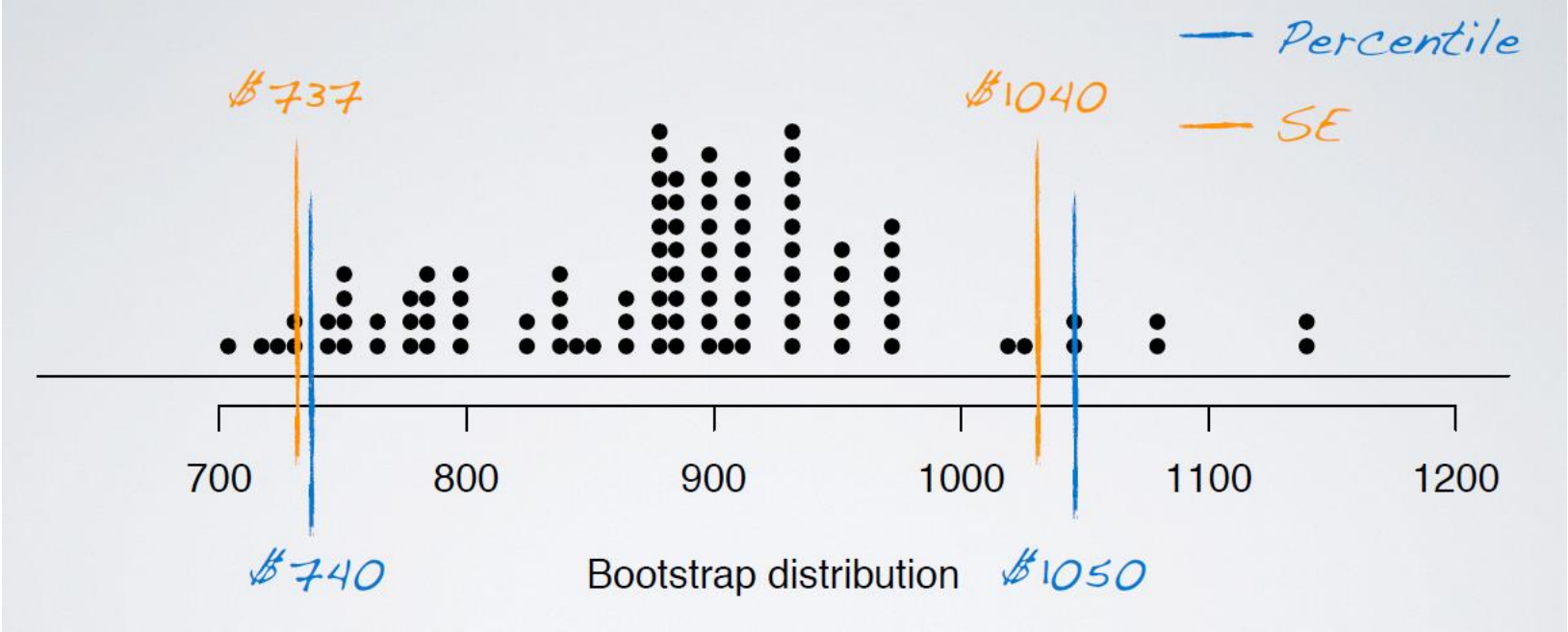
$= 882.515 \pm 1.65 \times 89.5758$

$\approx (734.7,\ 1030.3)$



Bootstrapping

- 90% boostrap przedział ufności dla 100 bootstrap próbek



Ograniczenia metody bootsrapping

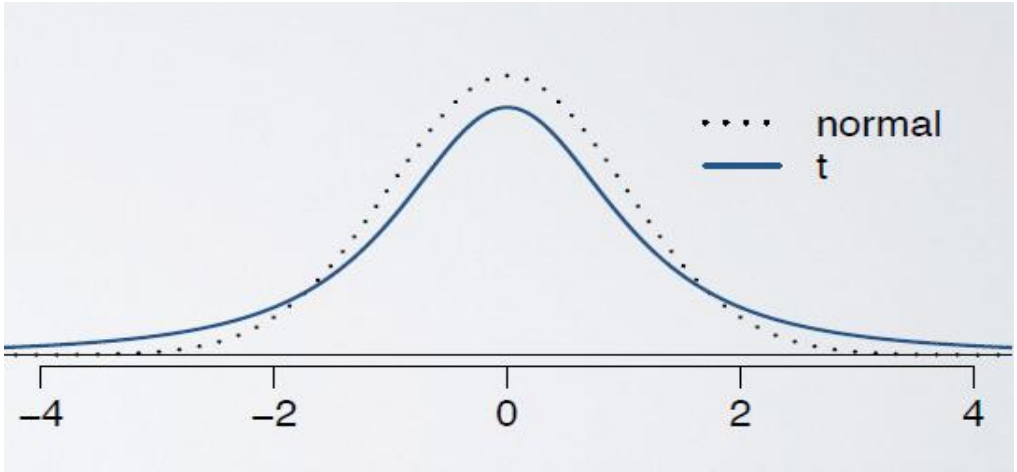
- Nie ma aż tak precyzyjnych warunków jak dla metody CLT
- Jeżeli rozkład pierwotny jest bardzo skrzywiony lub o małej statystyce metoda może nie być wiarygodna
- Pierwotna próbka musi być reprezentatywna, jeżeli ma bias to wynik będzie też miał ten bias.

Bootstrap vs próbka z populacji

- Próbka z populacji jest tworzona wybierając z całej populacji podzbiór (bez zwracania)
- Bootstrap zbiór jest tworzony wybierając z próbki elementy (ze zwracaniem)
- W obu przypadkach patrzymy na zmienną statystyczną takiej probki, np. średnią.

Duże czy małe próbki

- Jeżeli obserwacje są niezależne i rozkład danej zmiennej w populacji nie jest bardzo skrzywiony to
  - Rozkład średniej wielu próbek jest bliski do normalnego
  - Oszacowanie SE = σ/n jest wiarygodne
- Jeżeli próbka jest mała oraz σ jest nieznana to użyj t-rozkładu, podobny kształt ale większe ogony rozkładu.



t-rozkład

- Zawsze symetryczny względem 0 (jak standaryzowany rozkład normalny)
- Ma jeden parametr, **df** = ilość stopni swobody, które regulują jak duże są ogony rozkładu

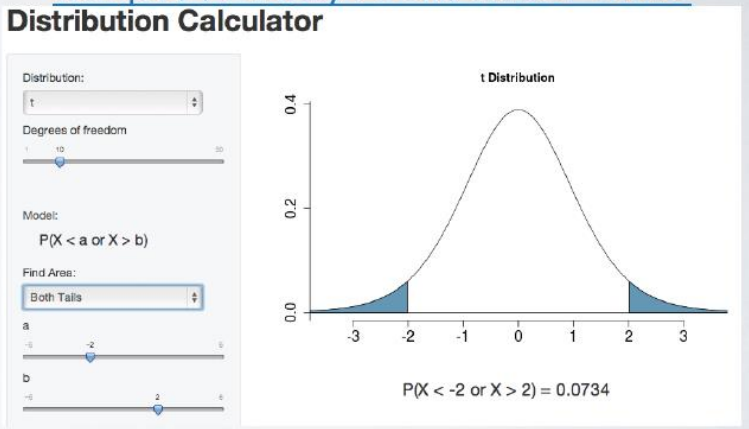
Jak zwiększamy ilość stopni swobody to rozkład zbliża się do normalnego

t- statystyka

- Dla wnioskowania statystycznego dla wartości średniej, kiedy:
  - Σ nieznane
  - $n < 30$
- Oblicz t-statystykę

$$T = \frac{obs - null}{SE}$$

http://bitly.com/dist_calc



- p-wartość (ta sama definicja co poprzednio)

Przykład

- Test statystyka = 2 oraz testujemy hipotezę H_0 dla dwustronnej granicy. Kiedy możemy odrzucić hipotezę na poziomie 95%CL?
  - Obliczmy p-wartość

a. $P(|Z| > 2)$ 0.0455 → reject

b. $P(|t_{df=50}| > 2)$ 0.0509 → fail to reject?

c. $P(|t_{df=10}| > 2)$ 0.0734 → fail to reject

Przykład

- Oszacowanie średniej dla małej próbki

$$\bar{x} \pm t^{\star}_{df} SE_{\bar{x}}$$

$$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^{\star}_{n-1} \frac{s}{\sqrt{n}}$$

$$df = n - 1$$

Przykład:

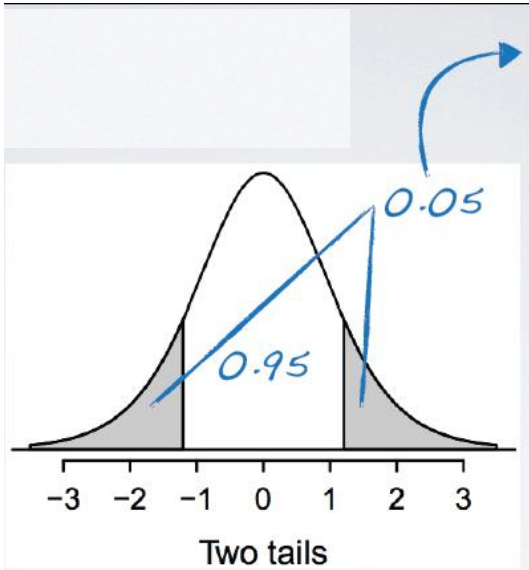
- Granie w gry komputerowe w trakcie obiadu powoduje zwiększenie spożycia ciasteczek w trakcie popołudnia

Zebrane dane:

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Przykład

- Policz df
- Wylicz zakres przedziału dla zadanego poziomu ufności



	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79
	26	1.31	1.71	2.06	2.48	2.78
	27	1.31	1.70	2.05	2.47	2.77

$df = 22 - 1 = 21$

R

```
> qt(0.025, df = 21)
[1] -2.079614
```

Przykład

- Na poziomie 95% ufności w oparciu o zebrane dane możemy stwierdzić że osoby które grały w gry w trakcie obiadu, spożywały później 32.1-72.1g ciasteczek.

$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$
$$t^{\star}_{21} = 2.08$$

$$\bar{x} \pm t^{*}\ SE = 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}}$$
$$= 52.1 \pm 2.08 \times 9.62$$
$$= 52.1 \pm 20 = (32.1, 72.1)$$

Przykład

- Czy te dane pozwalają aby stwierdzić że 30g byłoby niewystarczające?

$\bar{x} = 52.1\ g$

$s = 45.1\ g$

$n = 22$

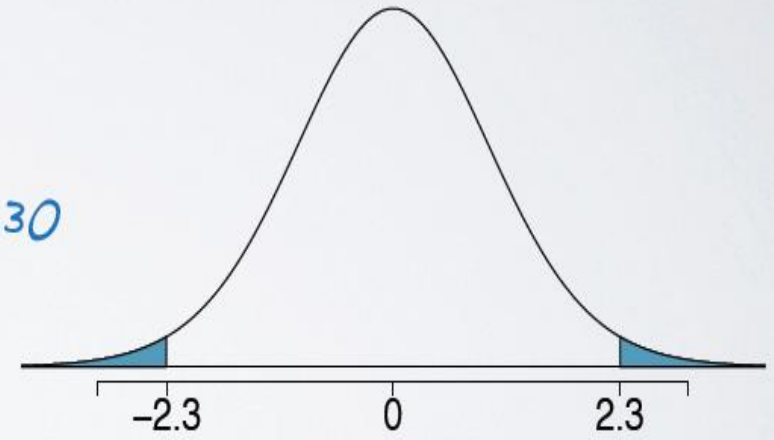
$SE = 9.62$

$H_0: \mu = 30$

$H_A: \mu \neq 30$

$$T = \frac{52.1 - 30}{9.62} = 2.30$$

$df = 22 - 1 = 21$



Przykład

- Czy te dane pozwalają aby stwierdzić że 30g byłoby niewystarczające?

$df = 21$

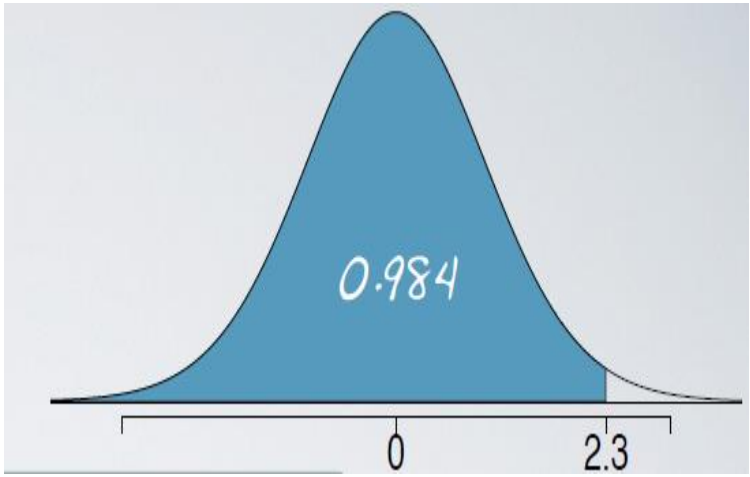
$0.02 < p\text{-value} < 0.05$

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77

Przykład

- Czy te dane pozwalają aby stwierdzić że 30g byłoby niewystarczające?

```
R
> pt(2.30, df = 21)
[1] 0.9840989
> 2 * pt(2.30, df = 21, lower.tail = FALSE)
[1] 0.03180228
```



95% confidence interval: (32.1 g, 72.1 g)

$H_0 : \mu = 30$

$H_A : \mu \neq 30$

p-value ≈ 0.0318

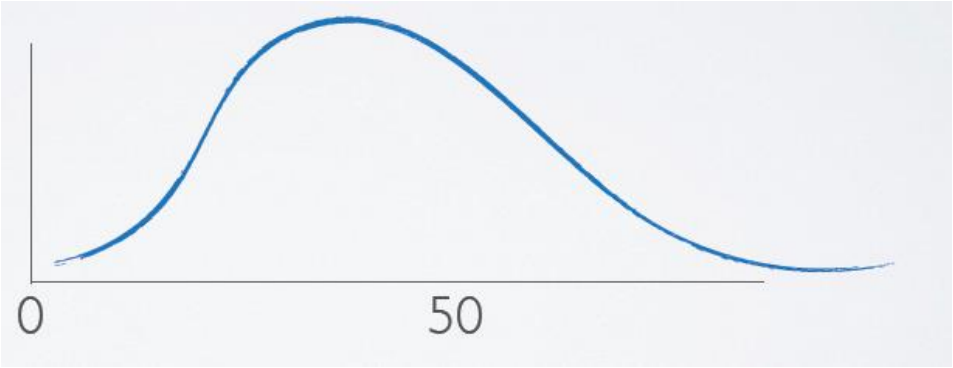
Reject H_0

agree

Przykład

- Sprawdzamy warunki:
  - Niezależność:
    - Losowy wybór osób
    - 22 < 10% wszystkich którzy grają w trakcie jedzenia
  - Skrzywienie rozkładu:

$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$



Przykład

- Porównujemy średnie dwóch małych próbek:

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

$$df = min(n_1 - 1, n_2 - 1)$$

confidence interval

point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm t^{\star}_{df} SE_{(\bar{x}_1 - \bar{x}_2)}$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

hypothesis test

$$T_{df} = \frac{obs - null}{SE}$$

$$T_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

Przykład

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

- Porównujemy średnie dwóch małych próbek:

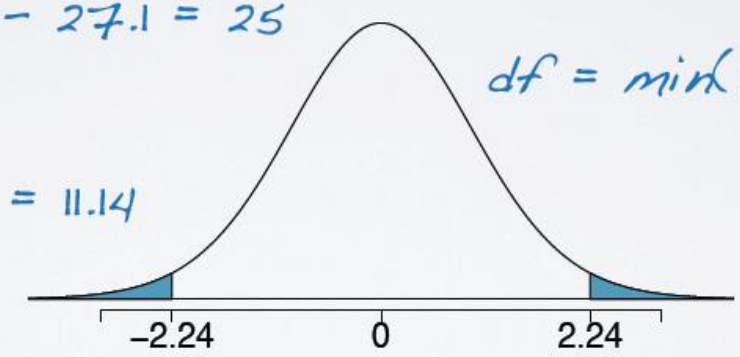
$H_0: \mu_{wd} - \mu_{wod} = 0 \quad H_A: \mu_{wd} - \mu_{wod} \neq 0$

$(\bar{x}_{wd} - \bar{x}_{wod}) = 52.1 - 27.1 = 25$

$SE = \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}} = 11.14$

$T = \frac{25 - 0}{11.14} = 2.24$

$df = \min(22 - 1, 22 - 1) = 21$



$(\bar{x}_{wd} - \bar{x}_{wod}) \pm t^* SE = 25 \pm 2.08 \times 11.14$

$= 25 \pm 23.17$

$= (1.83, 48.17)$

95% confidence interval: (1.83g, 48.17g)

$H_0 : \mu_{wd} - \mu_{wod} = 0$

$H_A : \mu_{wd} - \mu_{wod} \neq 0$

p-value ≈ 0.04

Reject H_0 — agree

Analiza wariancji

- Aby porównać „średnie” dla 3+ grup używamy nowego testu nazywanego „analizą wariancji” (ANOVA) i nowej statystyki nazywanej **F-statystyką**.
- H_0: wszystkie średnie mają ta samą wartość

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

- H_A: przynajmniej jedna średnia jest różna

Analiza wariancji

z/t test

- Porównaj średnie dla dwóch grup. Czy ich różnica może być wytłumaczona przez statystyczną fluktuację?

$$H_0 : \mu_1 = \mu_2$$

anova

- Porównaj średnie dla trzech lub więcej grup. Czy ich różnica może być wytłumaczona przez statystyczną fluktuację?

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

Analiza wariancji

z/t test

- Policz test statystykę: stosunek

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

anova

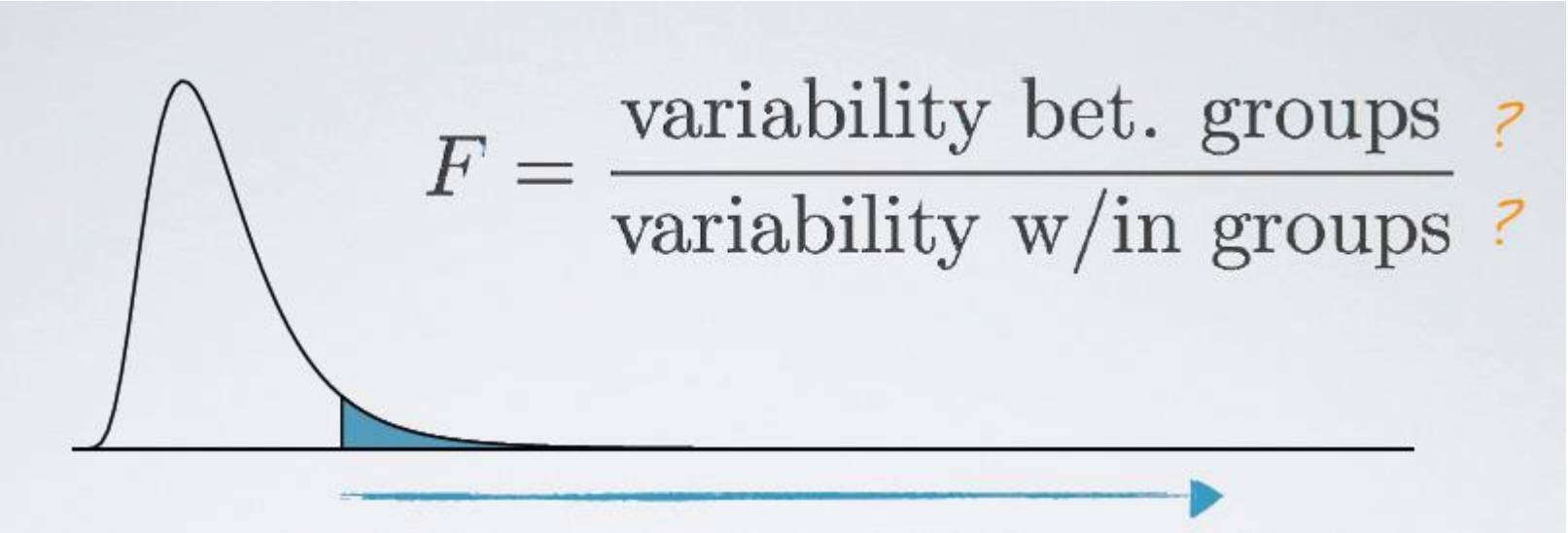
- Policz test statystykę: stosunek

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

Duża wartość statystyki testowej wskazuje na małe prawdopodobieństwo. Jeżeli prawdopodobieństwo bardzo małe odrzuć hipotezę H_0

Analiza wariancji

- Aby odrzucić H0, potrzebujemy uzyskać dużą wartość F-statystyki. Wymaga to aby zmienność pomiędzy grupami była dużo większa niż zmienność w ramach grup.



Przykład

score

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
4	5	working class
5	6	working class
6	6	working class
…	…	…
795	9	middle class

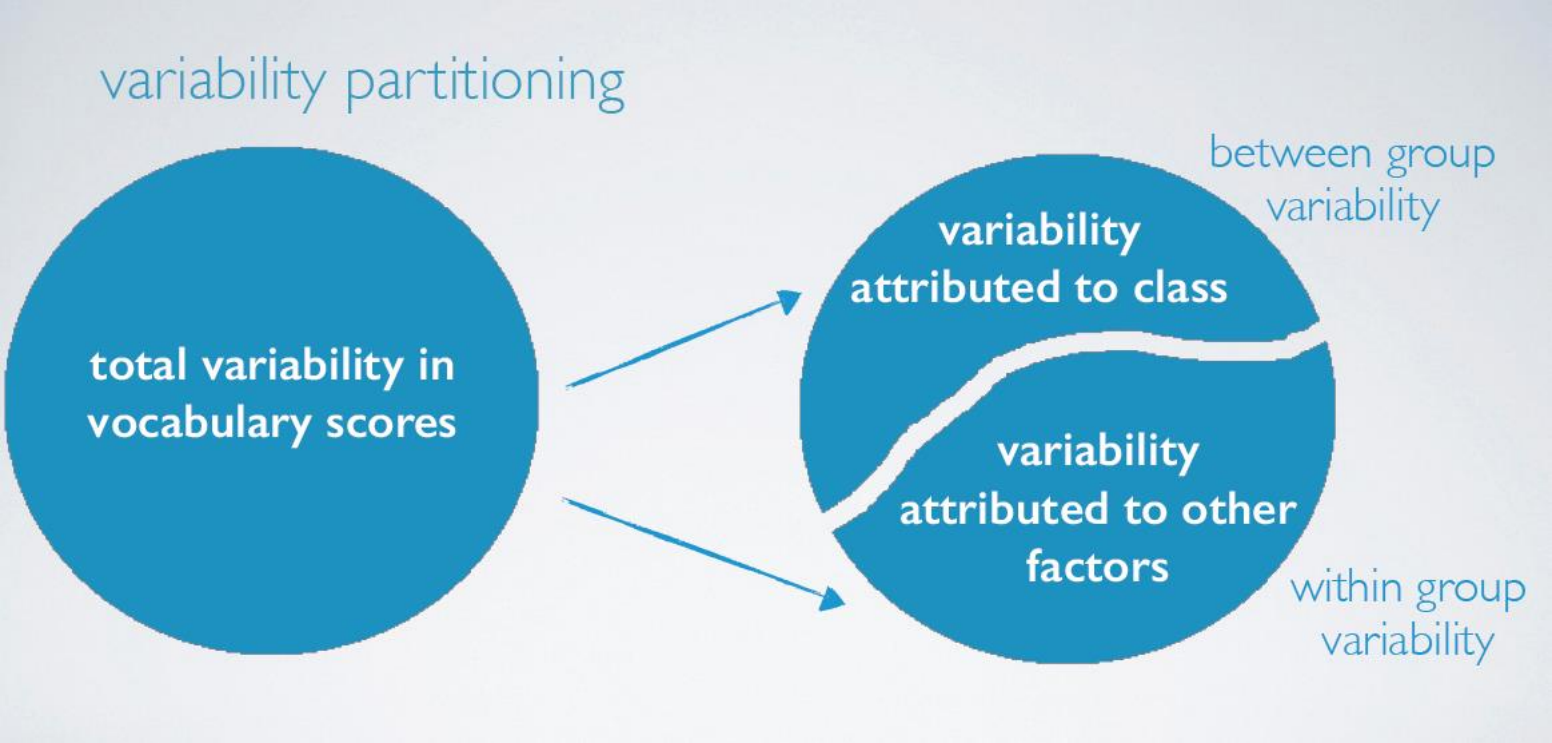
	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

H_0: The mean outcome is the same across all categories

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_A: At least one pair of means are different from each other

Analiza zmienności



SST

Sum of squares total (SST):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y_i : value of the response variable for each observation

$\bar{y}$: grand mean of the response variable

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
…	…	…
795	9	middle class

	n	mean	sd
overall	795	6.14	1.98

$$SST = (6-6.14)^2 + (9-6.14)^2 + (6-6.14)^2 + \cdots + (9-6.14)^2 = 3106.36$$

Oznacza zmienność całej próbki

SSG

Sum of squares group (SSG):

$$SSG = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

n_j : number of observations in group j
$\bar{y}_j$: mean of the response variable for group j
$\bar{y}$: grand mean of the response variable

	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

$$SSG = (41 \times (5.07 - 6.14)^2) + (407 \times (5.75 - 6.14)^2) + (331 \times (6.76 - 6.14)^2) + (16 \times (6.19 - 6.14)^2) \approx 236.56$$

Oznacza zmienność pomiędzy grupami: różnica pomiędzy średnią w grupie i średnią w całej próbce, ważona ilością elementów w grupie.

SSE

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.8			
	Total		3106.36			

Sum of squares error (SSE):

$$SSE = SST - SSG$$

Nie wyjaśniona zmienność

3106.36 - 236.56 = 2869.8

Teraz chcemy przejść do średnich wartości

degrees of freedom

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56			
Error	Residuals	791	2869.80			
	Total	794	3106.36			

Degrees of freedom associated with ANOVA:

- total: $df_T = n - 1$ → 795 − 1 = 794
- group: $df_G = k - 1$ → 4 − 1 = 3
- error: $df_E = df_T - df_G$ → 794 − 3 = 791

Teraz chcemy przejść do średnich wartości

mean square error

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855		
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

Mean squares: Average variability between and within groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom.

- group: $MSG = SSG/df_G$ → 236.56 / 3 ≈ 78.855
- error: $MSE = SSE/df_E$ → 2869.8 / 791 ≈ 3.628

F – statystyka

F statistic

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

F statistic: Ratio of the between group and within group variability:

$$F = \frac{MSG}{MSE}$$

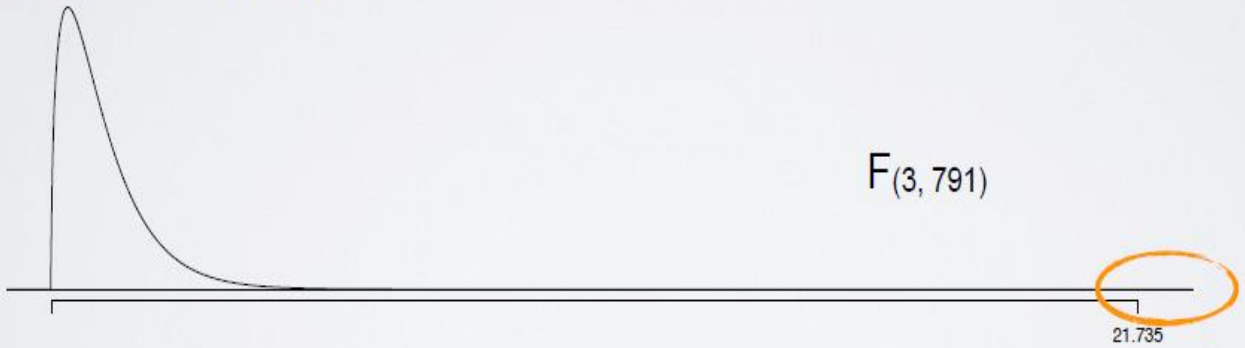
$$\frac{78.855}{3.628} \approx 21.735$$

p-wartość

p-value

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

- p-value is the probability of at least as large a ratio between the "between" and "within" group variabilities if in fact the means of all groups are equal
- area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



p-wartość

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

using R

```
R
> pf(21.735, 3, 791, lower.tail = FALSE)
[1] 1.559855e-13
```

using the applet

http://bitly.com/dist_calc

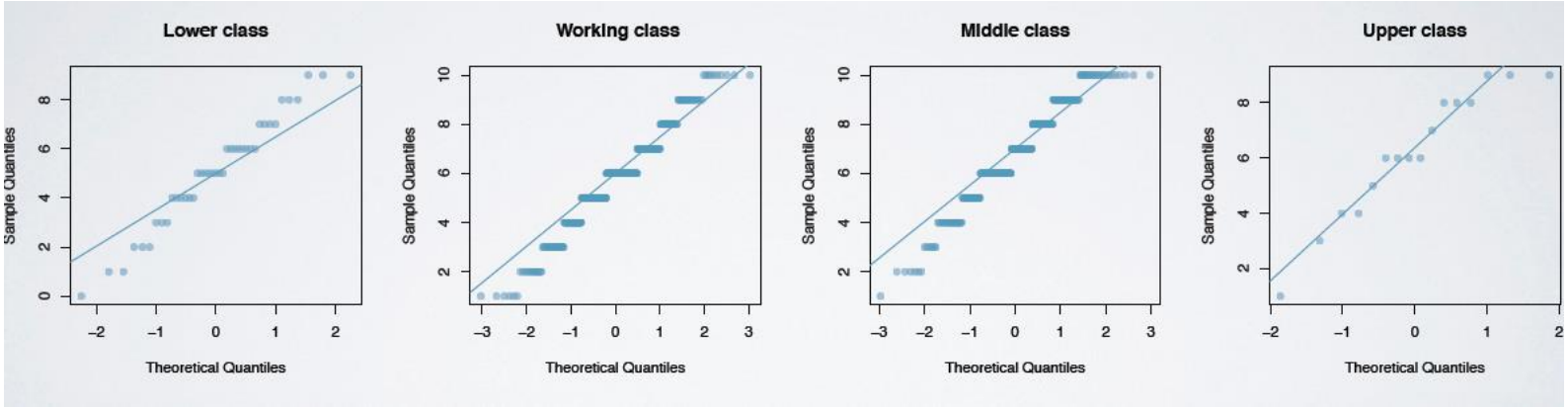
Odrzucamy hipotezę H_0

Kiedy możemy stosować analizę wariancji?

- **Niezależność**:
  - Obserwacje muszą być niezależne
  - Każda nj < 10% odpowiedniej grupy
  - Grupy muszą być od siebie niezależne (nie ma parowania)
- Rozkład obserwowanej wielkości powinien mieć **rozkład prawie normalny**, szczególnie ważne jeżeli próbki są małe; sprawdzamy przy pomocy rozkładu percentili

Kiedy możemy stosować analizę wariancji?

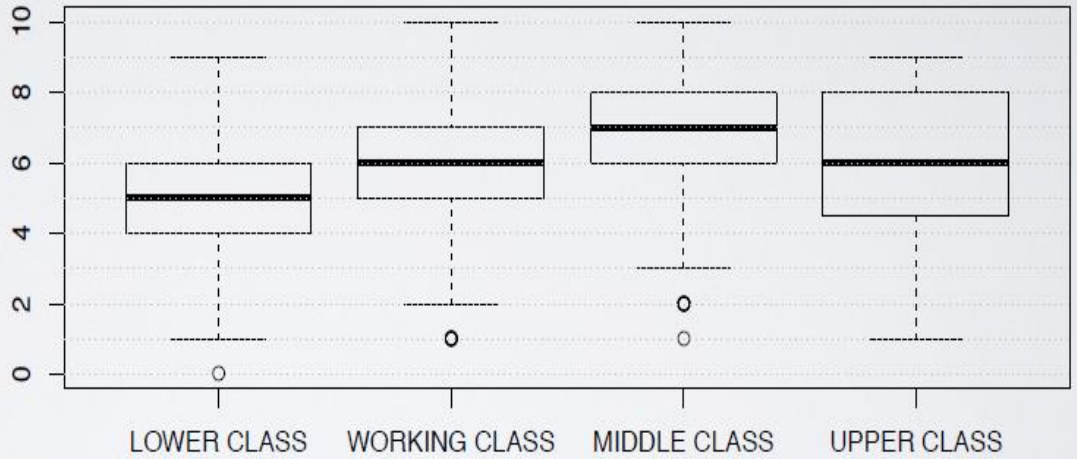
- Rozkład obserwowanej wielkości powinien mieć **rozkład prawie normalny**, szczególnie ważne jeżeli próbki są małe; sprawdzamy przy pomocy rozkładu percentili



Kiedy możemy stosować analizę wariancji?

- Zmienność obserwowanej wielkości powinna być porównywalna w każdej grupie

	n	sd
lower class	41	2.24
working class	407	1.87
middle class	331	1.89
upper class	16	2.34
overall	795	1.98



Porównywania wielokrotne

- Można testować różne pary grup
- Ale wtedy stosujemy skalowanie dla poziomu ufności: **Bonferroni poprawki**

Bonferroni correction:

$$\alpha^{\star} = \alpha / K \quad K: \text{number of comparisons}, K = \frac{k(k-1)}{2}$$

Przykład

- Wracamy do naszego przykładu: mamy 4 grupy, testujemy parami aby stwierdzic które grupy są znacząco różne

$$k = 4$$

$$K = \frac{4 \times 3}{2} = 6$$

$$a^* = 0.05 / 6 \approx 0.0083$$

Porównywanie parami

Używamy konsystentnej definicji błędu standardowego oraz ilości stopni swobody.

Standard error for multiple pairwise comparisons:

$$SE = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

indep. groups test:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Degrees of freedom for multiple pairwise comparisons:

$$df = df_E$$

$$df = min(n_1 - 1, n_2 - 1)$$

Przykład

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
class	3	236.56	78.855	21.735	<0.0001
Residuals	**791**	2869.80	**3.628**		
Total	794	3106.36			

$H_0: \mu_{middle} - \mu_{lower} = 0$

$H_A: \mu_{middle} - \mu_{lower} \neq 0$

	n	mean
lower class	41	5.07
middle class	331	6.76

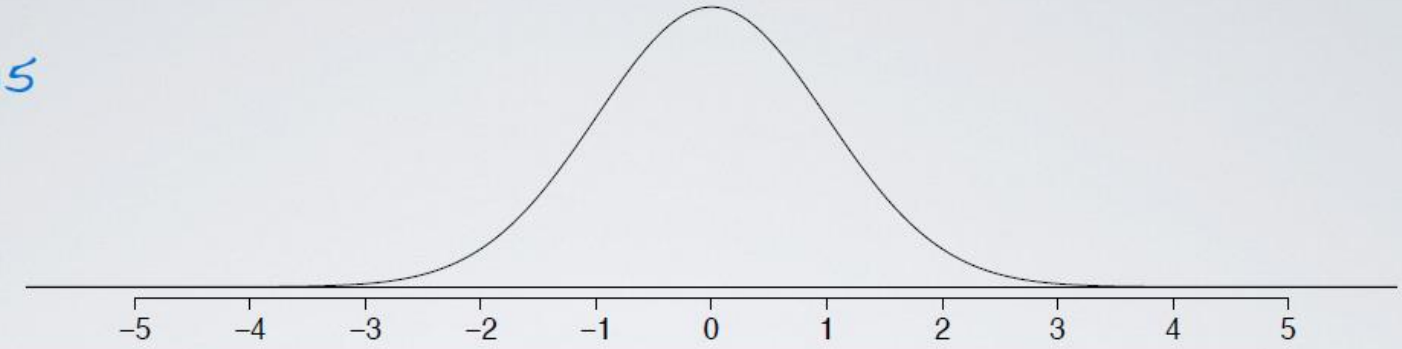
$$T = \frac{(\bar{X}_{middle} - \bar{X}_{lower}) - 0}{\sqrt{\frac{MSE}{n_{middle}} + \frac{MSE}{n_{lower}}}} = \frac{(6.76 - 5.07)}{\sqrt{\frac{3.628}{331} + \frac{3.628}{41}}} = \frac{1.69}{0.315} = 5.365$$

$df = 791$

Przykład

$T = 5.365$

$df = 791$



R

```
> 2 * pt(5.365, df = 791, lower.tail = FALSE)
[1] 1.063895e-07
```

$a^* = 0.0083$

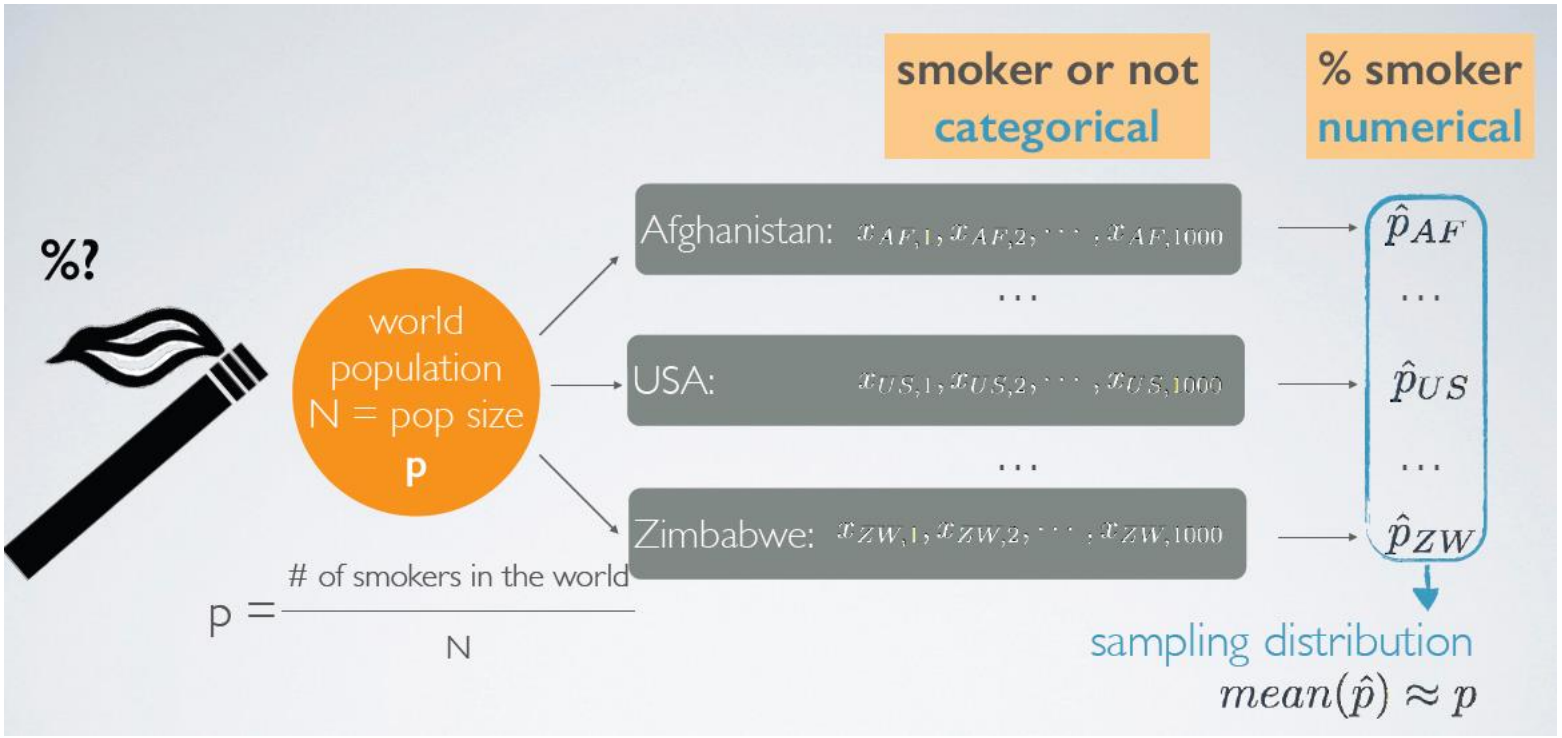
$p\text{-value} < a^* \rightarrow$ Reject H_0

Wnioskowanie statystyczne dla zmiennych opisowych

Definiowanie „proporcji”

Przykład

- Definiowanie „proporcji”



Centralne twierdzenie graniczne

Rozkład proporcji w próbce jest bliski do rozkładu normalnego, symetryczny względem proporcji dla całej populacji, błąd standardowy jest odwrotnie proporcjonalny do wielkości próbki

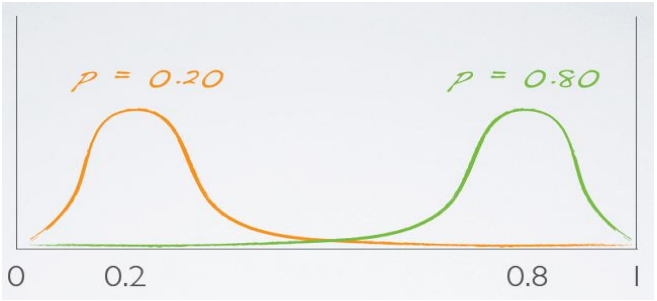
$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

shape center spread

Warunki:

- Niezależność obserwacji
- Rozmiar próbki/skrzywienie rozkładu: powinno być co najmniej

$np \geq 10$ and $n(1-p) \geq 10$.



Testowanie hipotezy dla pojedynczej proporcji

- Zdefiniuj hipotezę:

$$H_0 : p = null\ value$$
$$H_A : p < \ or \ > \ or \ \neq \ null\ value$$

- Policz jaka jest proporcja w badanej próbce
- Sprawdź warunki:
  - Niezależność
  - Rozmiar próbki
- Narysuj rozkład, policz t-statystykę

$$Z = \frac{\hat{p} - p}{SE}, \quad SE = \sqrt{\frac{p(1-p)}{n}}$$

- Podejmij decyzję
  - Jeżeli p-wartość < α, odrzuć H_0
  - Jeżeli p-wartość > α, nie możesz odrzucić H_0

Testowanie hipotezy dla pojedynczej proporcji

$\hat{p}$ vs. p

	confidence interval	hypothesis test
success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

Różnica między dwoma proporcjami

$$(\hat{p}_1 - \hat{p}_2) \pm z^\star SE_{(\hat{p}_1 - \hat{p}_2)}$$

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- ✓ $n_1p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
- ✓ $n_2p_2 \geq 10$ and $n_2(1-p_2) \geq 10$