

# ***TESTOWANIE ŁĄCZNEJ ISTOTNOŚCI PARAMETRÓW W KMNRL***

## ***(CZYLI O TEŚCIE F)***

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

# Plan wykładu

1) Wprowadzenie

2) Test  $F$

a. Układ hipotez statystycznych

b. Statystyka testowa

c. Wartość krytyczna i zbiór krytyczny

d. Konkluzja testu

e.  $P$ -value

3) Szczególny przypadek testu  $F$

4) Przykład

# Wprowadzenie

- Przypomnijmy punkt wyjścia naszych rozważań, czyli zapis równania regresji (w domyśle: spełniającej wszystkie założenia KMNRL):

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + \varepsilon_t$$

- Dotychczas, w ramach stosowania podstawowych technik wnioskowania statystycznego w obrębie KMNRL, uwagę skupialiśmy na:

- 1) testowaniu pojedynczego współczynnika regresji – za pomocą testu  $t$ -Studenta
- 2) budowie przedziałów ufności (blisko “spowinowaconych” z testem  $t$ ) – oczywiście, także dotyczących pojedynczych parametrów

- W tym materiale poznamy jedną z technik wnioskowania o kilku parametrach jednocześnie – choć ograniczymy się do bardzo konkretnego kontekstu, a mianowicie do testowania łącznej istotności (równoważnie: nieistotności) grupy parametrów. Innymi słowy, przedmiotem sprawdzenia będzie jednoczesne zerowanie się pewnego zestawu parametrów (to zaś których – określamy sami).

- Przykładowo, bazując na powyższym, ogólnym zapisie modelu, załóżmy, że interesuje nas testowanie łącznej istotności parametrów  $\beta_1$  i  $\beta_2$  (tylko). Stosowny układ hipotez wyglądałby następująco:

$$H_0: \beta_1 = 0 \wedge \beta_2 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0 \vee \beta_2 \neq 0$$

przy czym hipoteza alternatywna stanowi zwykłe zaprzeczenie hipotezy zerowej, w której występuje koniunkcja zdań:  $\beta_1 = 0$  i  $\beta_2 = 0$ . Zatem w  $H_1$  mamy alternatywę zaprzeczeń (= zaprzeczenie koniunkcji). Zwróćmy też uwagę, że  $H_0$  możemy równoważnie zapisać jako:  $H_0: \beta_1 = \beta_2 = 0$ .

## Wprowadzenie

- Naturalnie, możemy w tym miejscu zadać sobie pytanie (z lekkim zdziwieniem czy oburzeniem;): *A to nie wystarczyłoby przeprowadzić, po prostu, kilku osobnych testów  $t$  – badając istotność każdego z interesujących nas współczynników, i na ich podstawie wysnuć wniosek odnośnie jednoczesnego zerowania się tych parametrów?*
- Wiadoma odpowiedź: nie :)
- Czas na wyjaśnienia... Wyobraźmy sobie sytuację, w której w ramach pewnego oszacowanego modelu okazuje się, że dwa jego parametry (dajmy na to,  $\beta_1$  i  $\beta_2$ , stojące przy zmiennych  $x_{t1}$  i  $x_{t2}$ , odpowiednio) są *indywidualnie nieistotne*, tzn. dla każdego z nich test  $t$ -Studenta istotności nie odrzuca  $H_0: \beta_i = 0$  ( $i = 1, 2$ ). Można wtedy powiedzieć, że usunięcie z modelu zmiennej  $x_{t1}$  ALBO  $x_{t2}$  **nie** spowoduje znacznego, “statystycznie istotnego” spadku mocy wyjaśniającej modelu (czyli zarazem też “statystycznie istotnego” wzrostu sumy kwadratów reszt – bo mniej zmiennych objaśniających w modelu oznacza wyższą SSE). Wobec tego, nie tracimy zbyt dużo redukując wyjściowy o tylko JEDNĄ spośród wymienionych zmienną, a wzrost SSE można uznać za zanedbywalny. Jednak to, że obie zmienne są *indywidualnie* nieistotne **nie** uprawnia nas jeszcze do stwierdzenia, że moglibyśmy usunąć z modelu je obie na raz – *może* (choć też nie musi) się bowiem okazać, że *jednoczesne* usunięcie obu regresorów spowoduje już zbyt duży wzrost SSE (tym samym spadek mocy wyjaśniającej modelu), aby uznać go za zanedbywalny.
- Powyższy tok rozumowania uzasadnia konieczność zastosowania testu innego niż  $t$ -Studenta – do tego, by odpowiedzieć na trzy następujące, równoważne pytania: 1) Czy parametry  $\beta_1$  i  $\beta_2$  są *łącznie* statystycznie (nie)istotne?, 2) Czy zmienne  $x_{t1}$  i  $x_{t2}$  są *łącznie* statystycznie (nie)istotne? 3) Czy można zredukować rozważany model (*jednocześnie*) o zmienne  $x_{t1}$  i  $x_{t2}$ ?

## Test $F$ : układ hipotez

- Przechodząc do przedstawienia procedury testu  $F$  na możliwie najwyższym poziomie ogólności, przygotujmy wpierw stosowny grunt. Na początek przypomnijmy punkt wyjścia, czyli ogólny zapis równania regresji liniowej:
- zapis skalarny:  $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + \varepsilon_t$
  - zapis macierzowy:  $y = X\beta + \varepsilon$
- Wyobraźmy sobie, że te parametry (spośród  $\beta_0, \beta_1, \dots, \beta_K$ ), o których łącznej (nie)istotności chcemy wnioskować, ujmujemy w wektor  $\beta^{(0)}$  o wymiarach  $k_0 \times 1$  (zatem  $k_0$  = liczba testowanych parametrów). (**Ważne:** nie muszą to być jakieś kolejne parametry – mogą nas interesować np.  $\beta_2, \beta_5$  i  $\beta_8$ ). Pozostałe parametry gromadzimy w wektorze oznaczonym jako  $\beta^{(1)}$  (o wymiarach  $(k - k_0) \times 1$ ). Wraz z podziałem wektora  $\beta$ :

$$\beta = \begin{bmatrix} \beta^{(0)} \\ \beta^{(1)} \end{bmatrix}$$

dokonujemy także odpowiedniego podziału macierzy  $X$ :

$$\underbrace{X}_{T \times k} = \left[ \underbrace{X^{(0)}}_{T \times k_0} \quad \underbrace{X^{(1)}}_{T \times (k - k_0)} \right]$$

przy czym macierz  $X^{(0)}$  grupuje te kolumny z macierzy  $X$ , które odpowiadają parametrom w wektorze  $\beta^{(0)}$ , zaś macierz  $X^{(1)}$  – wszystkie pozostałe. Wobec powyższego, zapis macierzowy modelu regresji możemy teraz przedstawić w następujący sposób:

$$y = X\beta + \varepsilon = X^{(0)}\beta^{(0)} + X^{(1)}\beta^{(1)} + \varepsilon$$

## Test $F$ : układ hipotez

- Układ hipotez w teście  $F$  możemy – na tym możliwie najogólniejszym poziomie rozważań – zapisać tak:

$$H_0: \beta^{(0)} = 0 \quad \text{vs.} \quad H_1: \beta^{(0)} \neq 0$$

przy czym wartość “0” znajdującą się w powyższych zapisach należy rozumieć jako wektor zerowy o wymiarach  $k_0 \times 1$  (czyli  $k_0$ -elementowa kolumna zer)

- Zapis hipotezy zerowej:  $H_0: \beta^{(0)} = 0$  oznacza koniunkcję zerowania się poszczególnych parametrów, które są przedmiotem testu
- Zapis hipotezy alternatywnej:  $H_1: \beta^{(0)} \neq 0$  stanowi zwykłe zaprzeczenie  $H_0$ , ponieważ wektor  $\beta^{(0)}$  jest różny od wektora zerowego  $\Leftrightarrow$  co najmniej jedna jego współrzędna (czyli co najmniej jeden spośród parametrów w  $\beta^{(0)}$ ) jest różna od zera.
- O czym w istocie świadczą obydwie hipotezy?

| $H_0: \beta^{(0)} = 0$                                                                                                       | $H_1: \beta^{(0)} \neq 0$                                                                                                                  |
|------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Testowane parametry są <i>łącznie</i> statystycznie <u>nie</u> istotne                                                       | Testowane parametry są <i>łącznie</i> statystycznie istotne                                                                                |
| Zmienne objaśniające odpowiadające testowanym parametrom są <i>łącznie</i> statystycznie <u>nie</u> istotne                  | Zmienne objaśniające odpowiadające testowanym parametrom są <i>łącznie</i> statystycznie istotne                                           |
| Rozważany model regresji można zredukować ( <i>jednocześnie</i> ) o zmienne objaśniające odpowiadające testowanym parametrom | Rozważanego modelu regresji <u>nie</u> można zredukować ( <i>jednocześnie</i> ) o zmienne objaśniające odpowiadające testowanym parametrom |

## Test $F$ : układ hipotez

### ➤ Kwestie terminologiczne:

W związku z przedstawionym powyżej znaczeniem hipotez testu  $F$ , czasami określa się mianem testu redukcji modelu, przy czym model, którego redukcja podlega testowaniu określamy wtedy mianem modelu wyjściowego, modelu niezredukowanego, modelu pełnego (w mowie potocznej: także modelu “oryginalnego”).

### ➤ W powyższej konwencji:

- $H_0$  opowiada się za modelem zredukowanym (“mniejszym”)
- $H_1$  – “sprzeciwiając” się redukcji – opowiada się za modelem pełnym

### ➤ Zauważmy, że również na “zwykły” test $t$ -Studenta istotności *pojedynczego* parametru można spojrzeć jako test redukcji modelu – choć tylko o *jedną* zmienną objaśniającą. Co więcej – i na szczęście – można pokazać, że gdyby właśnie na potrzeby testowania redukcji modelu o jeden dany parametr przeprowadzić zarówno test $t$ , jak i omawiany w tym materiale test $F$ , wówczas konkluzje wynikające z obydwu procedur będą zawsze spójne. Oznacza to, że oba testy – w sytuacji testowania pojedynczego współczynnika regresji – są sobie równoważne (co nie oznacza jednak, że otrzymamy tę samą wartość statystyk testowych, bo mają one inne rozkłady). W praktyce jednak, gdy chodzi o testowanie tylko jednego parametru, bezdyskusyjnie wybieramy test $t$ , gdyż – jak sami Państwo zobaczycie poniżej – test $F$ jest bardziej żmudny do przeprowadzenia (tak więc, po co wytaczać działo p/pancerne na wróbla...:)

## Test $F$ : układ hipotez

### ➤ Uwaga praktyczna

W praktyce stujemy nieco prostszy w odbiorze, bardziej bezpośredni zapis hipotez niż to przedstawiono powyżej (te powyższe zapisy są wymuszone chęcią przedstawienia zagadnienia na możliwie najwyższym poziomie ogólności). Wyobraźmy sobie, że mamy do czynienia z następującym modelem (znaczenie zmiennych jest nieważne):

$$A_t = \beta_0 + \beta_1 B_t + \beta_2 C_t + \beta_3 D_t + \varepsilon_t$$

Chcąc tu wnioskować o łącznej istotności, przykładowo, parametrów  $\beta_1$  i  $\beta_3$  (czyli zarazem: o łącznej istotności zmiennych  $B_t$  i  $D_t$ ), układ hipotez testu  $F$  zapiszemy tak:

$$H_0: \beta_1 = \beta_3 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0 \vee \beta_3 \neq 0$$

Taki bezpośredni zapis jest bardziej czytelny aniżeli:

$$H_0: \beta^{(0)} = 0 \quad \text{vs.} \quad H_1: \beta^{(0)} \neq 0$$

gdzie  $\beta^{(0)} = \begin{bmatrix} \beta_1 \\ \beta_3 \end{bmatrix}$ .

→ Swoją drogą, zauważmy, że **błędny** byłby zapis  $H_1$  w “skrótowej” postaci:  $H_1: \beta_1 \neq \beta_3 \neq 0$  (bo kto powiedział, że parametry  $\beta_1$  i  $\beta_3$  mają być *między sobą* różne??!)

## Test $F$ : układ hipotez

### ➤ Jeszcze jedna uwaga praktyczna odnośnie zapisu hipotez

Zwróćmy uwagę, że w zapisie hipotez występują tylko parametry, nie zaś ich estymatory ( $\hat{\beta}_1, \hat{\beta}_3$ ) czy – uwaga(!) – zmienne objaśniające stojące przy nich. Pewnie jest to oczywiste, ale jednak (jak podpowiada doświadczenie) warto zauważyć, że zupełnie **błędny** byłby zapis “w stylu”:

$$H_0: B_t = D_t = 0 \quad \text{vs.} \quad H_1: B_t \neq 0 \vee D_t \neq 0$$

Wszak wnioskujemy o *wpływie* zmiennych (ujmowanym przez *parametry*), a nie o ich *wartościach*.

## Test $F$ : statystyka testowa

➤ Statystyka testu  $F$  przyjmuje postać:

$$F_{emp} = \frac{\frac{SSE_0 - SSE_1}{k_0}}{\frac{SSE_1}{T - k}} \quad | H_0 \sim F(k_0, T - k)$$

gdzie:

- $SSE_0$  jest sumą kwadratów reszt w modelu **zredukowanym** (czyli tym odpowiadającym  $H_0$ )
- $SSE_1$  jest sumą kwadratów reszt w modelu **wyjściowym** (czyli tym odpowiadającym  $H_1$ )
- $k_0$  to liczba parametrów, które widnieją w  $H_0$
- $T - k$  to ta “stara” (w sensie: wykorzystywana uprzednio w testach  $t$ ) liczba stopni swobody – czyli ta w modelu ogólnym, któremu odpowiada  $H_1$ 
  - Zauważmy, że wyrażenie  $\frac{SSE_1}{T - k}$  to nic innego, jak wariancja resztowa ( $s^2$ ) w modelu wyjściowym
- zapis  $F(a, b)$  oznacza tzw. **rozkład  $F$**  o pewnych parametrach (ogólnie)  $a$  i  $b$ , przy czym każdy z nich to **liczba stopni swobody** (pierwsza wynosząca  $a$ , zaś druga równa  $b$  – ich kolejność ma znaczenie!)

## Test $F$ : statystyka testowa

- Zatem do obliczenia wartości statystyki testowej potrzebujemy (m.in.)  $SSE_0$  i  $SSE_1$ . Zauważmy, że oszacowawszy pewien model “wyjściowy” (np. za pomocą REGLINP lub “na piechotę”) dysponujemy już wartością  $SSE_1$  (do tej pory “zwyczajnie” oznaczaną jako  $SSE$ , gdyż – z uwagi na brak alternatywnych modeli – nie było potrzeby indeksować tej wielkości). Natomiast chcąc przeprowadzić test  $F$  **potrzebujemy oszacować “na boku” model w wersji zredukowanej – o te zmienne, których parametry są zapisane w  $H_0$** . Z całej tej estymacji dodatkowego, “mniejszego” modelu, w zasadzie interesuje nas tylko suma kwadratów reszt, oznaczona właśnie symbolem  $SSE_0$ .
- Jeśli rozważyć wcześniej zarysowany **przykład** (trzy strony wyżej), to w celu obliczenia  $SSE_0$  należałoby pomocniczo (“na boku”) oszacować model postaci:

$$A_t = \beta_0 + \beta_2 C_t + \varepsilon_t$$

będący zredukowaną wersją modelu wyjściowego – zredukowaną o zmienne  $B_t$  i  $D_t$ , czyli te, których łączną istotność testujemy.

→ **Pytanie kontrolne:** Ile wynosiłaby w tym przypadku wartość  $k_0$ ?

→ **Przykład ten rozpracujemy dokładniej jeszcze na samym końcu tej prezentacji**

## Test F: statystyka testowa

- Można by tu zadać pytanie: Czy faktycznie trzeba ten zredukowany model SZACOWAĆ “od nowa”? Czy nie można by po prostu w miejsce jego parametrów wstawić ocen owych parametrów, ale tych uzyskanych już w modelu pełnym (i przy takowych obliczyć  $SSE_0$ )? Przecież mamy je już policzone, więc może by i dla modelu zredukowanego “stykły”? No, nie stykły by ;P A to dlatego, że te ostatnie (czyli oceny parametrów uzyskane w modelu “oryginalnym”) na pewno nie minimalizują sumy kwadratów reszt w modelu z innym (tu: mniejszym) zestawem regresorów. Choć owszem, jest to całkiem prawdopodobne, że oceny parametrów w modelu zredukowanym – te wynikające z jego faktycznego oszacowania “od nowa” za pomocą MNK – będą / mogą być dość podobne do tych uzyskanych w modelu wyjściowym. Niemniej jednak, będą inne :) Dlatego – powtórzmy – model zredukowany musimy faktycznie OSZACOWAĆ osobno / “od nowa”, nie występując się przy tym wyznaczonymi wcześniej ocenami parametrów w modelu ogólnym.

## Test $F$ : statystyka testowa

- Wróćmy do samej formuły na statystykę testową – zauważmy, że proste przekształcenia prowadzą nas do nieco mniej piętrowego zapisu statystyki, łatwiejszego do wykorzystania w praktyce:

$$F_{emp} = \left( \frac{SSE_0}{SSE_1} - 1 \right) \cdot \frac{T - k}{k_0} \mid H_0 \sim F(k_0, T - k)$$

- Zgodnie z powyższymi zapisami wzorów na statystykę testową, przy prawdziwości  $H_0$  (i, oczywiście, spełnieniu założeń KMNRL) statystyka  $F_{emp}$  ma **tw. rozkład  $F$**  (Fishera-Snedecora) o dwóch **liczbach stopni swobody**: pierwsza wynosząca  $k_0$ , zaś druga to  $T - k$ .

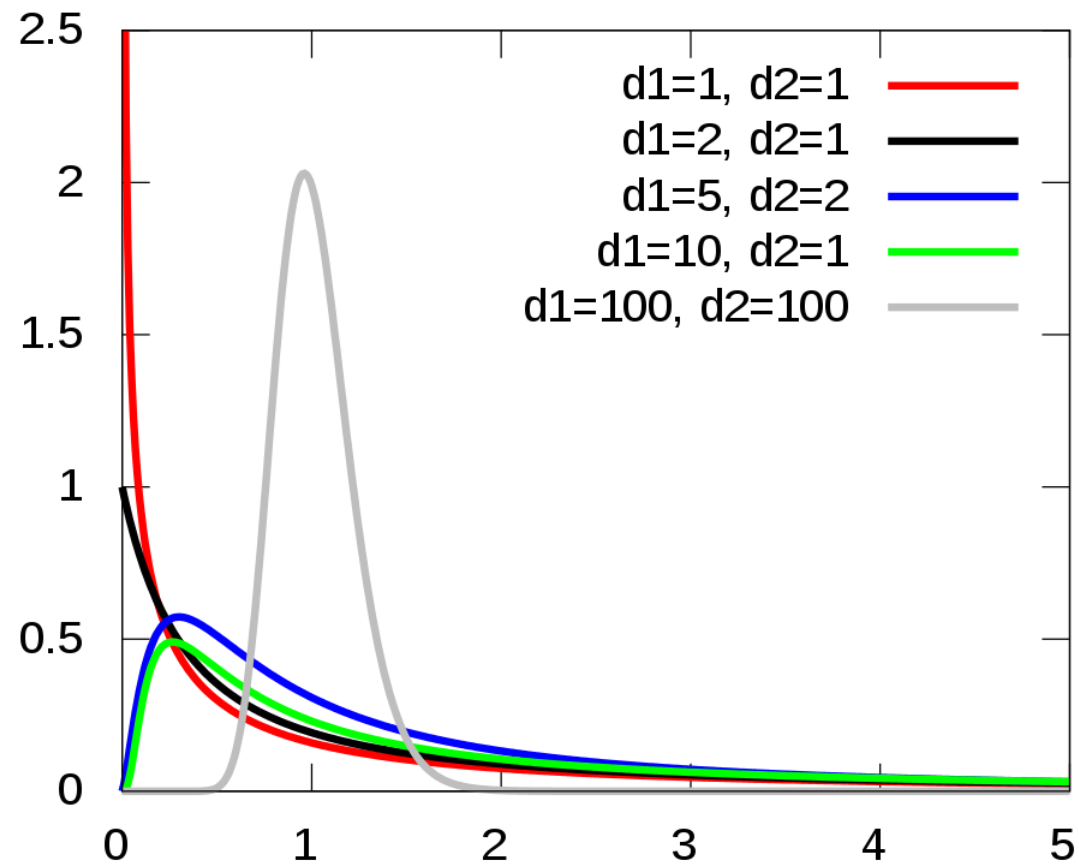
→ **Uwaga**: kolejność tych (jedynych) dwóch parametrów rozkładu  $F$  **nie** jest przemienna!!! Zatem rozkład np.  $F(2, 4)$  to **inny** rozkład aniżeli  $F(4, 2)$ .

→ Nazwa testu pochodzi od nazwy rozkładu, zaś ta ostatnia – od nazwiska pierwszego z dwóch jego odkrywców: Sir Ronalda Fishera.

- Powyższe stwierdzenie oznacza, że jeśli tylko  $H_0$  miałaby być prawdziwa, wówczas – przy założeniu, że moglibyśmy nieskończenie wiele razy generować próbę losową jako zbiór danych opisujących modelowane zjawisko – rozkład wartości  $F_{emp}$  otrzymanych dla tych wszystkich próbek pokrywa się z pewnym rozkładem teoretycznym, jakim jest tu (dobrze znany w nauce) rozkład  $F$ , a dokładniej  $F(k_0, T - k)$ . Jak zatem wygląda taki rozkład (jego funkcja gęstości)? Patrz następna strona...

## Test $F$ : statystyka testowa

- Wykresy funkcji gęstości przykładowych rozkładów  $F$  (stopnie swobody oznaczone jako “d1”, “d2”)



Źródło: <https://en.wikipedia.org/wiki/F-distribution> (dostęp: 20.04.2020)

## Test $F$ : statystyka testowa

➤ Spoglądając na wykres umieszczony na poprzedniej stronie, zauważamy, że rozkład  $F$  jest:

1) określony tylko na **półosi dodatniej** (tj.  $\mathbb{R}_+$ )

2) prawostronnie asymetryczny, z **wygasającym do zera prawym ogonem**

➤ **Pierwsza własność** oznacza w praktyce, że statystyka  $F_{emp}$  może przyjmować wartości tylko **dodatnie**

→ Przyjrzyjmy się formule wzoru na  $F_{emp}$ :

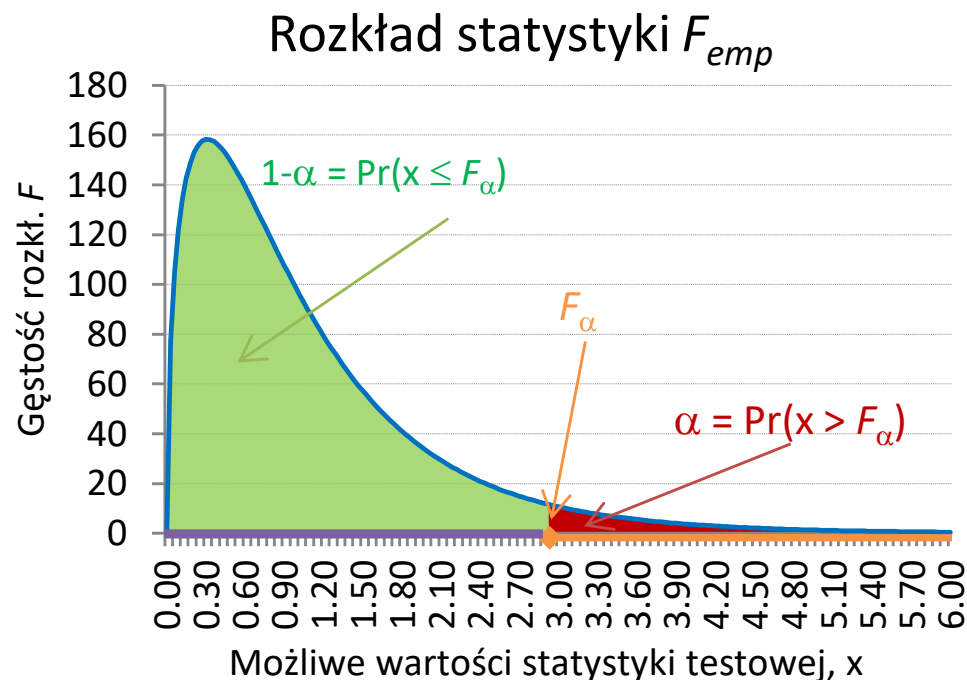
$$F_{emp} = \frac{\frac{SSE_0 - SSE_1}{k_0}}{\frac{SSE_1}{T - k}}$$

Zauważmy, że znak powyższego wyrażenia zależy tylko od znaku różnicy  $SSE_0 - SSE_1$ , ponieważ wszystkie pozostałe wielkości są dodatnie. Jednak różnica ta także przyjmuje tylko wartości dodatnie. Dlaczego?  $SSE_0$  to suma kwadratów reszt w modelu zredukowanym (o zmienne odpowiadające parametrom w  $\beta^{(0)}$ ), czyli w modelu “mniejszym” (z mniejszą liczbą parametrów), a zatem  $SSE_0$  musi być (mniej lub bardziej ale) większa od  $SSE_1$  (w modelu “większym” udaje się wyjaśnić większą część zmienności obserwacji, zatem mniej pozostaje zmienności niewyjaśnionej, ujmowanej właśnie przez  $SSE_1$ ). W związku z tym:  $SSE_0 - SSE_1 > 0$ , co oznacza, że także  $F_{emp} > 0$  (zawsze!)

➤ **Druga własność** od razu wskazuje, gdzie znajdować się będą mało wiarygodne (w świetle  $H_0$ ) wartości statystyki testowej, tj. te charakteryzujące się niskim prawdopodobieństwem ich odnotowania, jeśli  $H_0$  miałyby być prawdziwa – w prawym ogonie rozkładu. Przechodzimy zatem do... (następny slajd :)

## Test $F$ : wartość krytyczna i zbiór krytyczny

- Przyjrzyjmy się poniższemu wykresowi, przedstawiającemu rozkład  $F(3, 30)$  (liczby stopni swobody zostały tu arbitralnie dobrane, dla przykładu)



- $\alpha \cdot 100\%$ -wą **wartość krytyczną** (czyli na poziomie istotności  $\alpha$ ) w teście  $F$  oznaczamy symbolem  $F_\alpha$ 
  - W Excelu obliczamy ją za pomocą funkcji:  
 $\text{=ROZKŁAD.F.ODW}(\text{prawdopodobieństwo}; \text{stopnie\_swobody1}; \text{stopnie\_swobody2})$   
w miejsce argumentów wstawiając kolejno:  $(\alpha; k_0; T - k)$ , zachowując kolejność stopni swobody
- **Zbiór krytyczny** to przedział  $(F_\alpha, +\infty)$

## Test $F$ : konkluzja testu

### ➤ Reguła decyzyjna:

- Jeżeli  $F_{emp} \in Z_k$  (równoważnie:  $F_{emp} > F_\alpha$ ), wówczas **odrzucaamy**  $H_0$  na rzecz  $H_1$ .

Wnioski (wszystkie trzy są równoważne):

- test odrzuca redukcję modelu, zatem redukcja ta **nie** jest zasadna
- test wskazuje na **łącną** (statystyczną) **istotność** testowanych parametrów, a tym samym...
- test wskazuje na **łącną** (statystyczną) **istotność** testowanych zmiennych objaśniających

- Jeżeli  $F_{emp} \notin Z_k$  (równoważnie:  $F_{emp} \leq F_\alpha$ ), wówczas **nie mamy podstaw do odrzucenia**  $H_0$ .

Wnioski (wszystkie trzy są równoważne):

- test **nie** odrzuca redukcji modelu, zatem redukcja ta wydaje się zasadna
- test wskazuje na **łącną** (statystyczną) **nieistotność** testowanych parametrów, a tym samym...
- test wskazuje na **łącną** (statystyczną) **nieistotność** testowanych zmiennych objaśniających

### ➤ Dodatkowa refleksja:

Na podstawie wykresu na poprzedniej stronie możemy zauważyć, że to odpowiednio wysokie wartości statystyki testowej testu  $F$  będą świadczyć przeciwko  $H_0$  i, tym samym, na rzecz  $H_1$ . **Pytanie:** który komponent wzoru na  $F_{emp}$  jest głównie za to odpowiedzialny? **Odpowiedź:** Jest nim różnica pomiędzy sumami kwadratów reszt w modelu zredukowanym ( $SSE_0$ ) i w tym wyjściowym ( $SSE_1$ ) – im większa, tym większe szanse na odrzucenie  $H_0$ , co też jest bardzo intuicyjne...

## Test $F$ : $p$ -value

- **Wartość  $p$**  w teście  $F$  obliczamy w Excelu za pomocą formuły:

=ROZKŁAD.F( $x$ ; *stopnie\_swobody1*; *stopnie\_swobody2*)

przy czym w miejsce argumentu  $x$  podajemy obliczoną wartość statystyki testowej,  $F_{emp}$

- **Wnioskowanie na podstawie  $p$ -value** przebiega tu według tych samych reguł, co w przypadku testu  $t$

## Test F: szczególny przypadek

- Przypomnijmy, że **podstawowy schemat** (taki “szablon”) estymacji i analizy modelu regresji liniowej obejmuje kilka elementów:
  - Oceny punktowe parametrów strukturalnych
  - Błędy średnie szacunku
  - Mierniki dopasowania
  - Testy *t*-Studenta istotności poszczególnych parametrów strukturalnych
  - Przedziały ufności dla poszczególnych parametrów strukturalnych
  - ...a także – o czym nie było wcześniej mowy:
    - Test łącznej istotności wszystkich współczynników regresji (z wyjątkiem wyrazu wolnego)
- W ramach tego ostatniego badacz chce po prostu i od razu po estymacji modelu dowiedzieć się, czy wszystkie uwzględnione w modelu zmienne objaśniające są łącznie istotne, czy też nie (i wtedy cały model jest w zasadzie od razu “do wyrzucenia”, bo zasadniczo, całym rozważanym zbiorem regresorów nie udało się wyjaśnić zmienności obserwacji choćby w najmniejszej, ale jednak statystycznie istotnej mierze).
- Test ten nie obejmuje (nie “rusza”) wyrazu wolnego – nigdy, ponieważ rola wyrazu wolnego (który, oczywiście, też jest parametrem strukturalnym) jest zgoła inna niż pozostałych współczynników, i zawsze powinien być obecny w modelu regresji, nawet w sytuacji, gdyby jako taki okazywał się statystycznie nieistotny. Podyktowane jest to pewnymi względami technicznymi – jeśli zaniedbać obecność wyrazu wolnego, wówczas oceny pozostałych parametrów będą obciążone, czyli albo przeszacowane, albo niedoszacowane. Ponadto, jego obecność jest też wymagana do zasadności innych, bardziej zaawansowanych testów diagnostycznych przeprowadzanych w modelach regresji.

## Test $F$ : szczególny przypadek

- W zarysowanej wyżej sytuacji, stosowny **układ hipotez** możemy zapisać w następujący sposób (rozważając ogólny zapis równania regresji:  $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + \varepsilon_t$ ):

$$H_0: \beta_1 = \beta_2 = \dots = \beta_K = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \dots \vee \beta_K \neq 0$$

→ Hipotezie zerowej odpowiada model (maksymalnie) zredukowany – do postaci:  $y_t = \beta_0 + \varepsilon_t$

→ Hipotezie alternatywnej odpowiada model wyjściowy

- Co do postaci **statystyki testowej**, okazuje się, że w tym wariancie testu  $F$  przyjmuje ona szczególną postać (można to udowodnić):

$$F_{emp} = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{T-k}} \mid H_0 \sim F(k-1, T-k)$$

lub – w postaci mniej “pagodowej”, a przez to łatwiejszej do (poprawnego) wprowadzenia w arkuszu:

$$F_{emp} = \frac{R^2}{1-R^2} \cdot \frac{T-k}{k-1} \mid H_0 \sim F(k-1, T-k)$$

→ Wszystkie wielkości w powyższych formułach (tj. współczynnik determinacji i obie liczby stopni swobody) “pochodzą” z **modelu wyjściowego**.

→ W związku z powyższym – uwaga! – na potrzeby przeprowadzenia tego szczególnego przypadku testu  $F$  **nie** ma potrzeby pomocniczego szacowania “na boku” modelu zredukowanego. Uff...

## Test $F$ : szczególny przypadek

➤ **Krótką refleksja odnośnie formuły na  $F_{emp}$ :** 
$$F_{emp} = \frac{R^2}{1-R^2} \cdot \frac{T-k}{k-1}$$

Także w rozważanym tu, szczególnym przypadku testu  $F$  to dostatecznie wysokie wartości statystyki testowej prowadzą do odrzucenia  $H_0$ , a tym samym – przyjęcia  $H_1$  (świadczącej o tym, że obecny w danym modelu zbiór *wszystkich zmiennych objaśniających* jest łącznie statystycznie istotny).

**Pytanie:** Kiedy wartości  $F_{emp}$  będą wysokie?

**Odpowiedź:** Przyjrzyjmy się przytoczonej wyżej formule – składa się ona z dwóch ilorazów połączonych iloczynem. Pierwszy czynnik to iloraz “dopasowania” ( $R^2$ ) do “niedopasowania” ( $1 - R^2 = \varphi^2$ ) – im większe dopasowanie (jednocześnie idące w parze z mniejszym niedopasowaniem), tym wartość tego ilorazu (a przez to i statystyki testowej) jest wyższa, co może prowadzić do odrzucenia  $H_0$ . Ma to całkiem intuicyjny sens: odpowiednio wysokie dopasowanie zawdzięczamy wszakże obecności jakichś “sensownych” regresorów w modelu, więc są one wszystkie łącznie – być może z jakimiś pozostałymi “bezsensownymi” zmiennymi – statystycznie istotne. Wiemy jednak, że wartość samego współczynnika  $R^2$  wzrasta nawet na skutek uwzględnienia w modelu “bezsensownych” zmiennych – wzrost ten, co prawda, nie będzie wtedy tak “pokaźny” (istotny), jak przy włączeniu do modelu jakichś “zasadnych” zmiennych objaśniających, ale – będzie. Test  $F$  okazuje się w jakiś sposób przeciwdziałać sytuacji, w której pierwszy czynnik w  $F_{emp}$  będzie wysoki na skutek “zbyt dużej” liczby zmiennych (przez co podejrzewamy, że te zmienne są “śmieciowe”) – odpowiedzialny jest za to drugi iloraz:  $\frac{T-k}{k-1}$ . Przy zadanej liczbie obserwacji ( $T$ ) jego wartość zależy tylko od liczby regresorów (które to przyczyniają się do takiego, a nie innego dopasowania, a zatem też i wartości pierwszego czynnika we wzorze na  $F_{emp}$ ). Zauważmy, że im wyższa liczba regresorów ( $k - 1$ , nie licząc wyrazu wolnego), tym wartość całego ilorazu  $\frac{T-k}{k-1}$  jest niższa. W szczególności, gdy liczba szacowanych parametrów (bez wyrazu wolnego) przekracza połowę liczby obserwacji (tj.  $k - 1 > 0,5T$ ), wówczas wyrażenie to staje się ułamkiem właściwym (tj. liczbą z przedziału  $(0, 1)$ ), który “tłumi” wartość pierwszego czynnika w formule  $F_{emp}$ .

## Test $F$ : szczególny przypadek

➤ Czy REGLINP się tu przyda? :)

→ Wartość statystyki testowej  $F_{emp}$  dla rozważanej tu szczególnej postaci testu  $F$  jest dostępna w wynikach zwracanych przez funkcję REGLINP – proszę już samodzielnie sprawdzić, w którym dokładnie miejscu :)

→ Niemniej jednak, funkcja REGLINP nie “załatwia za nas całej sprawy” – to my potrzebujemy wiedzieć, ile wynoszą odpowiednie liczby stopni swobody, a także wyznaczyć wartość krytyczną czy  $p$ -value, i na ich podstawie sformułować konkluzję testu.

## Przykład

- Na koniec przeanalizujemy pewien przykład “na sucho” (bez żadnych konkretnych danych), aby uporządkować sobie kwestie operacyjne
- Rozważmy przykładowe równanie regresji, które już wcześniej pojawiło się w tej prezentacji (znaczenie poszczególnych zmiennych jest tu nieważne):

$$A_t = \beta_0 + \beta_1 B_t + \beta_2 C_t + \beta_3 D_t + \varepsilon_t$$

Jak widzimy, liczba parametrów strukturalnych (wszystkich, łącznie z wyrazem wolnym) wynosi tu  $k = 4$ . Ponadto, dajmy na to, że w wyniku oszacowania powyższego równania (powiedzmy, na podstawie  $T = 20$  obserwacji) uzyskano następujące rezultaty: suma kwadratów reszt = 10, współczynnik determinacji = 0,85.

W tym przykładzie ograniczamy się tylko do prześledzenia sposobu postępowania w rozwiązaniu dwóch zadań:

- 1) Przetestowanie łącznej istotności wszystkich (z wyjątkiem wyrazu wolnego) współczynników modelu (czyli jednakowoż: wszystkich zmiennych objaśniających)
- 2) Przetestowanie łącznej istotności parametrów  $\beta_1$  i  $\beta_3$  (tym samym: łącznej istotności zmiennych  $B_t$  i  $D_t$ , czyli *de facto* przedmiotem testu jest tu zasadność redukcji modelu wyjściowego o te dwie zmienne)

## Przykład

➤ **Ad 1)** Przetestowanie łącznej istotności wszystkich (z wyjątkiem wyrazu wolnego) współczynników tej regresji (czyli jednakowoż: wszystkich zmiennych objaśniających)

- **Układ hipotez:**  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$  vs.  $H_1: \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0$
- **Statystyka testowa** – zauważamy, że mamy tu do czynienia ze szczególnym przypadkiem testu  $F$ , więc sięgamy po stosowną formułę:

$$F_{emp} = \frac{R^2}{1-R^2} \cdot \frac{T-k}{k-1} \mid H_0 \sim F(k-1, T-k), \quad \text{przy czym: } k-1 = 3, T-k = 20-4 = 16$$

- **Obliczamy jej wartość**, mając na uwadze to, że nie potrzebujemy w tym przypadku pomocniczo, “na boku” szacować modelu zredukowanego (do trywialnej postaci  $A_t = \beta_0 + \varepsilon_t$ ), tylko korzystamy z wyników “już otrzymanych” dla modelu wyjściowego (a te mamy podane w treści przykładu):

$$F_{emp} = \frac{0,85}{1-0,85} \cdot \frac{20-4}{4-1} \approx 30,222$$

- **Wartość krytyczna** (przyjmujemy domyślny poziom istotności  $\alpha = 0,05$ ) – korzystając w Excelu z funkcji: =ROZKŁAD.F.ODW(0,05; 3; 16)  $\rightarrow F_\alpha = 3,239$
- **Zbiór krytyczny:**  $Z_k = (3,239; +\infty)$
- **Konkluzja:** Zauważamy, że  $F_{emp} \in Z_k$  (gdyż  $F_{emp} > F_\alpha$ ), zatem na poziomie istotności  $\alpha = 0,05$  odrzucamy  $H_0$  na rzecz  $H_1$ . Wobec tego stwierdzamy, że wszystkie parametry rozważanego modelu regresji (z pominięciem wyrazu wolnego) – a tym samym wszystkie zmienne objaśniające – są łącznie statystycznie istotne.

## Przykład

- Policzmy jeszcze ***p-value*** (celem sprawdzenia wrażliwości otrzymanej konkluzji na zmiany poziomu istotności:  $\text{=ROZKŁAD.F}(30,222; 3; 16) \rightarrow p\text{-value} = 7,96E - 07$

W notacji naukowej (ang. *scientific notation*) otrzymana wartość oznacza, że pierwsza cyfra znacząca (wynosząca 7) pojawia się dopiero na siódmym miejscu po przecinku. Zatem *p-value* jest tu bardzo niskie, znacząco (o kilka rzędów wielkości) niższe od każdego konwencjonalnego poziomu istotności, więc  $H_0$  jest odrzucana nie tylko przy  $\alpha = 0,05$ , ale także przy każdej innej typowej wartości poziomu istotności.

Wobec powyższego możemy sformułować następujący:

➤ **Wniosek/konkluzja (dot. *p-value*):**

Otrzymana wartość *p-value* jest niższa od każdego konwencjonalnego poziomu istotności, co oznacza, że hipoteza zerowa jest odrzucana przy każdej jego typowej wartości, nie tylko przy  $\alpha = 0,05$  (jak wskazano powyżej). Co więcej, *p-value* jest znacząco (bo aż o kilka rzędów wielkości) mniejsze od typowych wartości poziomu istotności, co skłania nas ku stwierdzeniu, że dane mocno sprzeciwiają się hipotezie zerowej, tym samym implikując znaczną łączną istotność wszystkich zmiennych objaśniających rozważanego modelu.

## Przykład

➤ **Ad 2)** Przetestowanie łącznej istotności parametrów  $\beta_1$  i  $\beta_3$  (tym samym: łącznej istotności zmiennych  $B_t$  i  $D_t$ , czyli *de facto* przedmiotem testu jest tu zasadność redukcji modelu wyjściowego o te dwie zmienne)

- **Układ hipotez:**  $H_0: \beta_1 = \beta_3 = 0$  vs.  $H_1: \beta_1 \neq 0 \vee \beta_3 \neq 0$

→ W  $H_0$  występują dwa parametry, zatem  $k_0 = 2$

- **Statystyka testowa** – zauważamy, że **nie** mamy tu do czynienia ze szczególnym przypadkiem testu  $F$ , więc sięgamy po stosowną, ogólną formułę:

$$F_{emp} = \left( \frac{SSE_0}{SSE_1} - 1 \right) \cdot \frac{T-k}{k_0} \mid H_0 \sim F(k_0, T-k), \text{ przy czym: } k_0 = 2, T-k = 20 - 4 = 16$$

- **Obliczamy jej wartość** – w tym celu potrzebujemy wartości **sum kwadratów reszt** w obydwu modelach:

→ **wyjściowym** – tę już mamy (jest podana w treści przykładu, jako już obliczona przy okazji estymacji głównego modelu): oznaczamy ją jako  $SSE_1 = 10$

→ **zredukowanym** – tym odpowiadającym  $H_0$ , czyli w modelu postaci:  $A_t = \beta_0 + \beta_2 C_t + \varepsilon_t$

Takiego modelu nie mamy, więc trzeba go sobie oszacować „na boku” → patrz kolejna strona

## Przykład

W celu oszacowania modelu zredukowanego (do postaci  $A_t = \beta_0 + \beta_2 C_t + \varepsilon_t$ ) potrzebujemy zapisać go macierzowo:

$$y = X^{(1)}\beta^{(1)} + \varepsilon$$

gdzie  $\beta^{(1)} = \begin{bmatrix} \beta_0 \\ \beta_2 \end{bmatrix}$ , zaś  $X^{(1)}$  stanowi „macierz  $X$ ” dla tego zredukowanego modelu. Dlaczego nie oznaczamy jej jako  $X^{(0)}$ ? Ponieważ zgodnie z ogólnymi zapisami przedstawionymi w tej prezentacji (w omówieniu układu hipotez testu), macierz oznaczona symbolem  $X^{(0)}$  zestawia wartości zmiennych objaśniających, ALE tych, o które model wyjściowy jest redukowany (czyli tu:  $B_t$  i  $D_t$ ), a nie tych, które pozostają w modelu zredukowanym (czyli tu: „1” i  $C_t$ ).

Tak więc na macierz  $X^{(1)}$  zmiennych objaśniających w modelu zredukowanym składają się  $k - k_0 = 2$  kolumny: pierwsza to kolumna „1” (odpowiadająca parametrowi  $\beta_0$ ), a druga – grupująca wartości zmiennej  $C_t$  (stojącej przy  $\beta_2$ ).

Zauważmy, że kolumna wartości zmiennej objaśnianej,  $y$ , jest dokładnie tą samą, co w modelu wyjściowym (wszak chcemy wyjaśnić dokładnie to samo zjawisko, tyle tylko, że pytamy, czy obejdzie się o „mniejszym” modelu).

Mając macierze  $y$  (ta sama, co w modelu wyjściowym) oraz  $X^{(1)}$  (utworzonej „na boku”) możemy oszacować model zredukowany (za pomocą MNK), wykorzystując funkcję REGLINP (też „na boku”), z której interesuje nas tylko suma kwadratów reszt – tę oznaczmy jako  $SSE_0$ . Wartość ta powinna wyjść większa niż  $SSE_1$  (z przyczyn, które wyjaśniliśmy już wcześniej). Dajmy na to, że  $SSE_0 = 12$ .

(Swoją drogą, warto podkreślić, że oceny parametrów  $\beta_0$  i  $\beta_2$  będą tu inne niż w modelu pełnym)

## Przykład

- Mając obliczoną wartość  $SSE_0$ , możemy wrócić do obliczenia wartości statystyki testowej:

$$F_{emp} = \left( \frac{SSE_0}{SSE_1} - 1 \right) \cdot \frac{T - k}{k_0} = \left( \frac{12}{10} - 1 \right) \cdot \frac{20 - 4}{2} = 1,6$$

- **Wartość krytyczna** (przyjmujemy domyślny poziom istotności  $\alpha = 0,05$ ) – korzystając w Excelu z funkcji: =ROZKŁAD.F.ODW(0,05; 2; 16)  $\rightarrow F_\alpha = 3,634$
- **Zbiór krytyczny:**  $Z_k = (3,634; +\infty)$
- **Konkluzja:** Zauważamy, że  $F_{emp} \notin Z_k$  (gdyż  $F_{emp} \leq F_\alpha$ ), zatem na poziomie istotności  $\alpha = 0,05$  nie mamy podstaw do odrzucenia  $H_0$ . Wobec tego stwierdzamy, że parametry  $\beta_1$  i  $\beta_3$  (a tym samym zmienne:  $B_t$  i  $D_t$ ) w rozważanym modelu regresji **nie** są łącznie statystycznie istotne. Tym samym model wyjściowy można by zredukować do postaci  $A_t = \beta_0 + \beta_2 C_t + \varepsilon_t$
- Policzmy jeszcze **p-value** (celem sprawdzenia wrażliwości otrzymanej konkluzji na zmiany poziomu istotności: =ROZKŁAD.F(1,6; 2; 16)  $\rightarrow p - value = 0,233$
- **Wniosek/konkluzja (dot. p-value):**  
Otrzymana wartość *p-value* jest wyższa od każdego konwencjonalnego poziomu istotności, co oznacza, że hipoteza zerowa **nie** jest odrzucana przy każdej jego typowej wartości, nie tylko przy  $\alpha = 0,05$  (jak wskazano powyżej).

The end :)