Wykład z analizy danych: kilka rozkładów i twierdzeń statystycznych

Marek Kubiak

Instytut Informatyki
Politechnika Poznańska

Cel wykładu

- poznanie kilku ważnych statystyk i rozkładów prawdopodobieństwa, które pojawiają się we wnioskowaniu statystycznym
- poznanie twierdzeń, które pokazują, w jakich sytuacjach wnioskowania statystycznego w sposób naturalny te statystyki i rozkłady będą się pojawiały (przy jakich hipotezach H_0)

Rozkład χ^2 (chi kwadrat)

Niech $X_1, X_2, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o jednakowych rozkładach $N(0, 1)$. Wtedy o zmiennej losowej:

$$Y = X_1^2 + X_2^2 + \ldots + X_n^2$$

mówimy, że ma ciągły **rozkład chi kwadrat** z n stopniami swobody, co oznacza się jako:

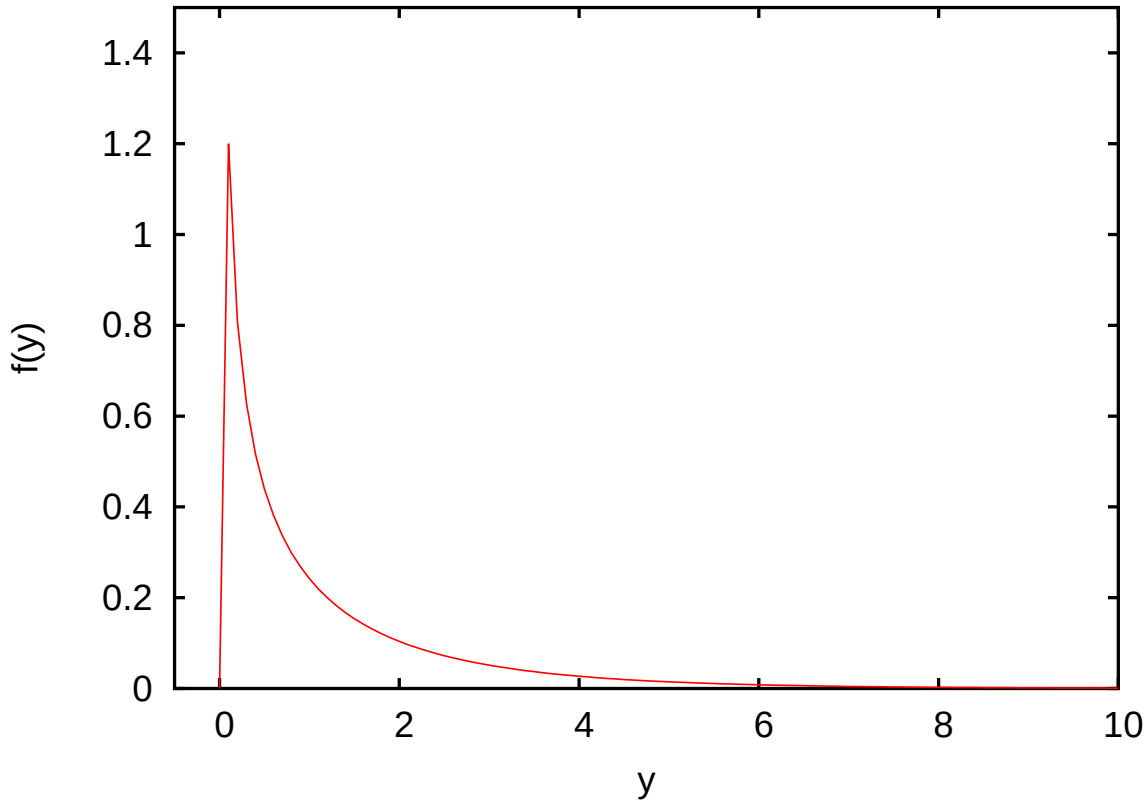
$$Y \sim \chi^2(n)$$

Rozkład χ^2 (chi kwadrat)

Komentarz do definicji

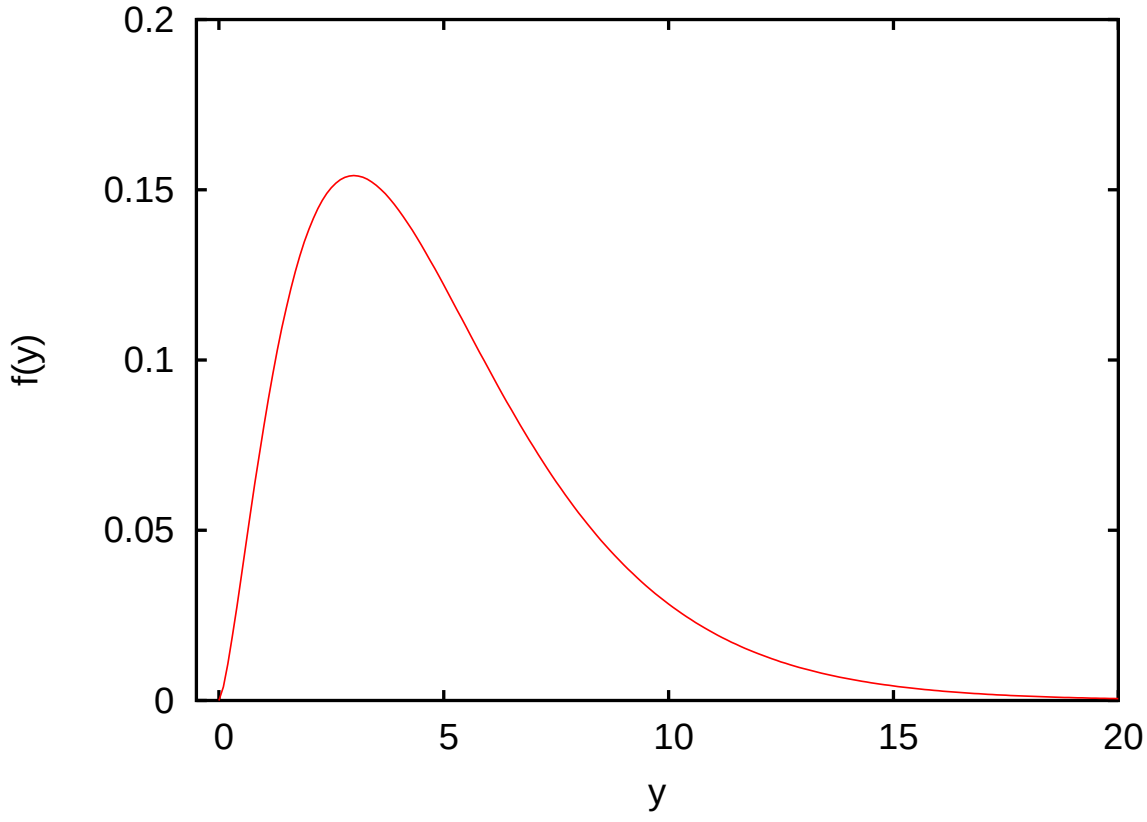
- zmienna o rozkładzie chi kwadrat ma tylko wartości nieujemne (suma kwadratów)
- rozkład chi kwadrat powstaje przy sumowaniu kwadratów niezależnych odchyleń o tej samej skali (pojawia się w wielu kontekstach)
- liczba stopni swobody pochodzi od liczby niezależnych zmiennych X_i w definicji
- wartości prawdopodobieństw dla tego rozkładu są w tablicach

Rozkład χ^2 (chi kwadrat)



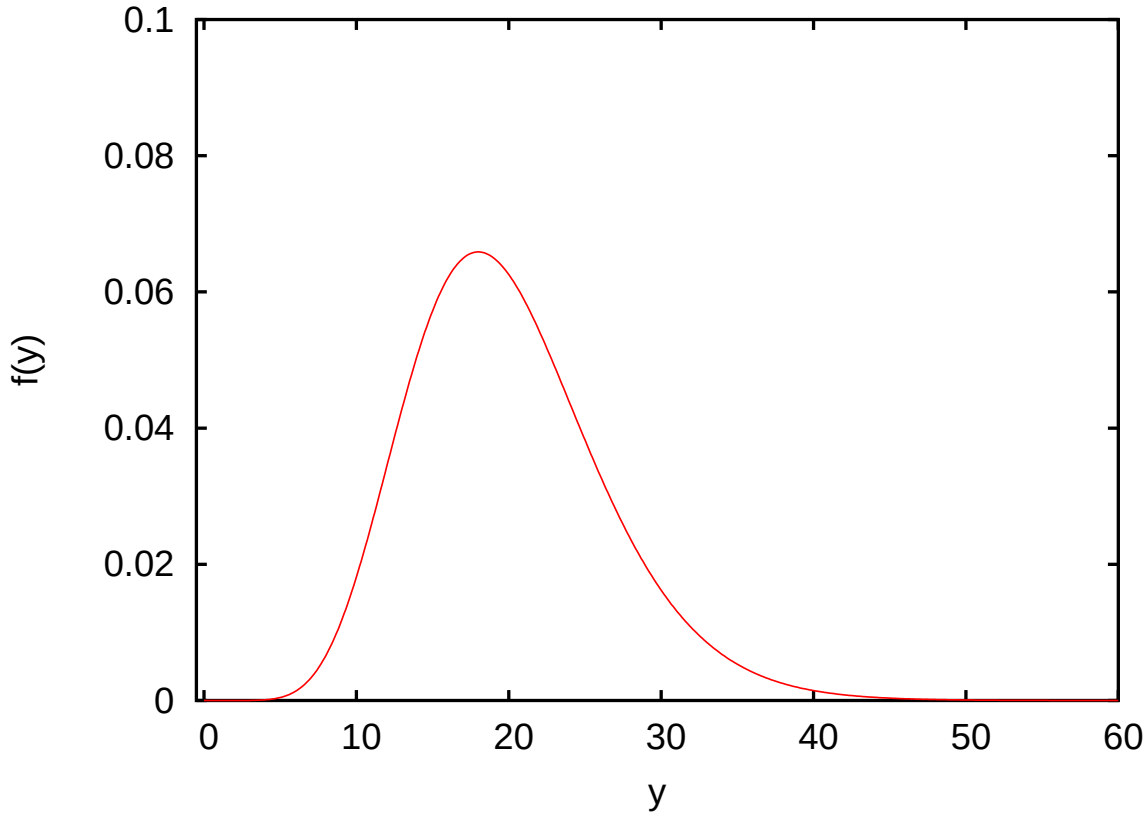
Rysunek: Rozkład prawdopodobieństwa zmiennej $Y \sim \chi^2(1)$

Rozkład χ^2 (chi kwadrat)



Rysunek: Rozkład prawdopodobieństwa zmiennej $Y \sim \chi^2(5)$

Rozkład χ^2 (chi kwadrat)



Rysunek: Rozkład prawdopodobieństwa zmiennej $Y \sim \chi^2(20)$

Twierdzenie Fishera

Niech $X_1, X_2, \ldots, X_n$ $(n > 1)$ będą niezależnymi zmiennymi losowymi o jednakowym rozkładzie $N(\mu, \sigma)$, $\sigma > 0$. Wtedy zmienne losowe:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

są niezależne, oraz mają rozkłady:

$$U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1) \qquad \frac{(n-1)S^{*2}}{\sigma^2} \sim \chi^2(n-1)$$

Twierdzenie Fishera

Komentarz do twierdzenia

- właśnie na podstawie tego twierdzenia jest zbudowany test na μ w rozkładzie normalnym – wiadomo, jaki będzie rozkład zmiennej U przy założeniu prawdziwości hipotezy $H_0 : \mu = \mu_0$ i przy znanej wariancji σ^2
- liczność próby dla testu na μ w rozkładzie normalnym nie musi być duża; to twierdzenie nie jest graniczne, ale prawdziwe także dla małych n
- rozkład estymatora wariancji z próby z rozkładu normalnego to chi kwadrat (przyda się to do testu na wariancję)
- liczba stopni swobody estymatora wariancji z próby to $n - 1$; jedna zmienna jest ustalona przez oszacowanie z próby wartości $\overline{X}$

Rozkład t (Studenta)

Niech X, Y będą niezależnymi zmiennymi losowymi o rozkładach:

$$X \sim N(0,1) \qquad Y \sim \chi^2(n)$$

Wtedy o zmiennej losowej:

$$t = \frac{X}{\sqrt{\frac{1}{n}Y}}$$

mówimy, że ma **rozkład t Studenta** o n stopniach swobody, co oznacza się jako:

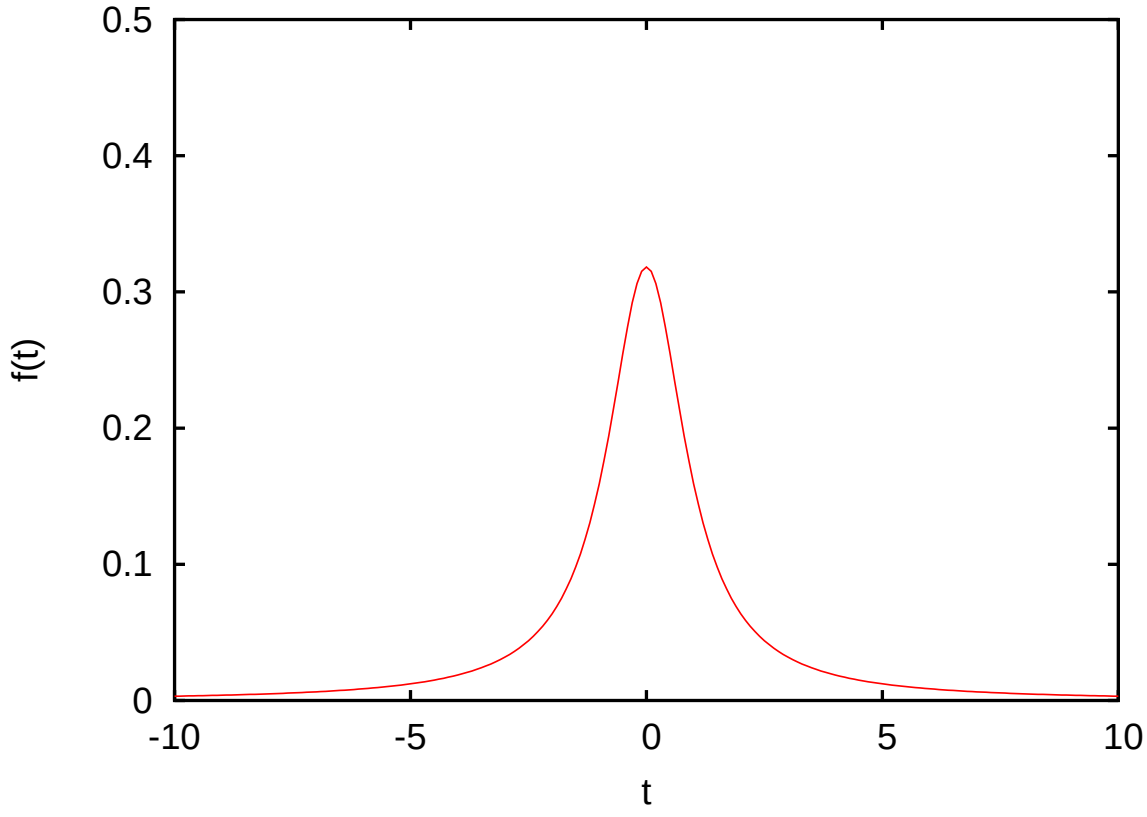
$$t \sim t(n)$$

Rozkład t (Studenta)

Komentarz do definicji

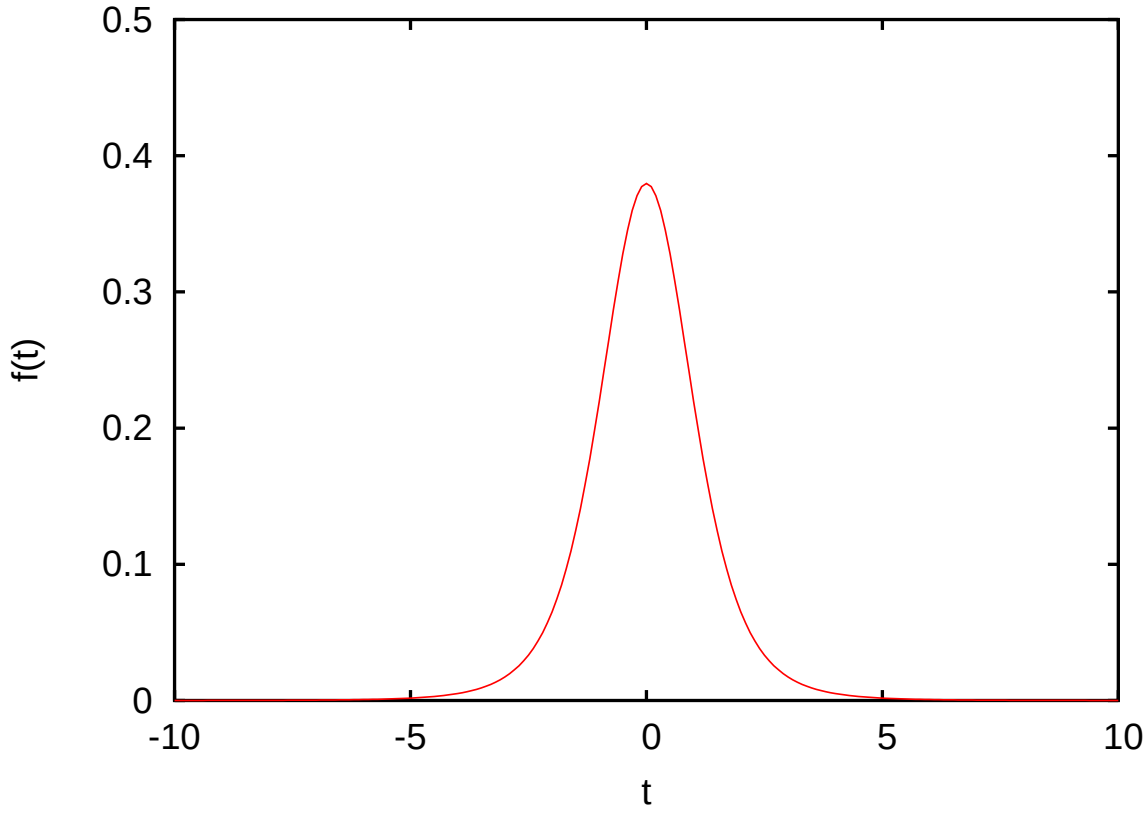
- zmienna o rozkładzie t ma wartości dowolne
- rozkład t Studenta powstaje przy dzieleniu średniej z próby przez prosto przekształcony estymator wariancji z próby
- liczba stopni swobody rozkładu t pochodzi od liczby stopni swobody rozkładu χ^2
- wartości prawdopodobieństw dla rozkładu t są w tablicach
- dla dużych n rozkład t Studenta zbiega się do rozkładu $N(0, 1)$

Rozkład t (Studenta)



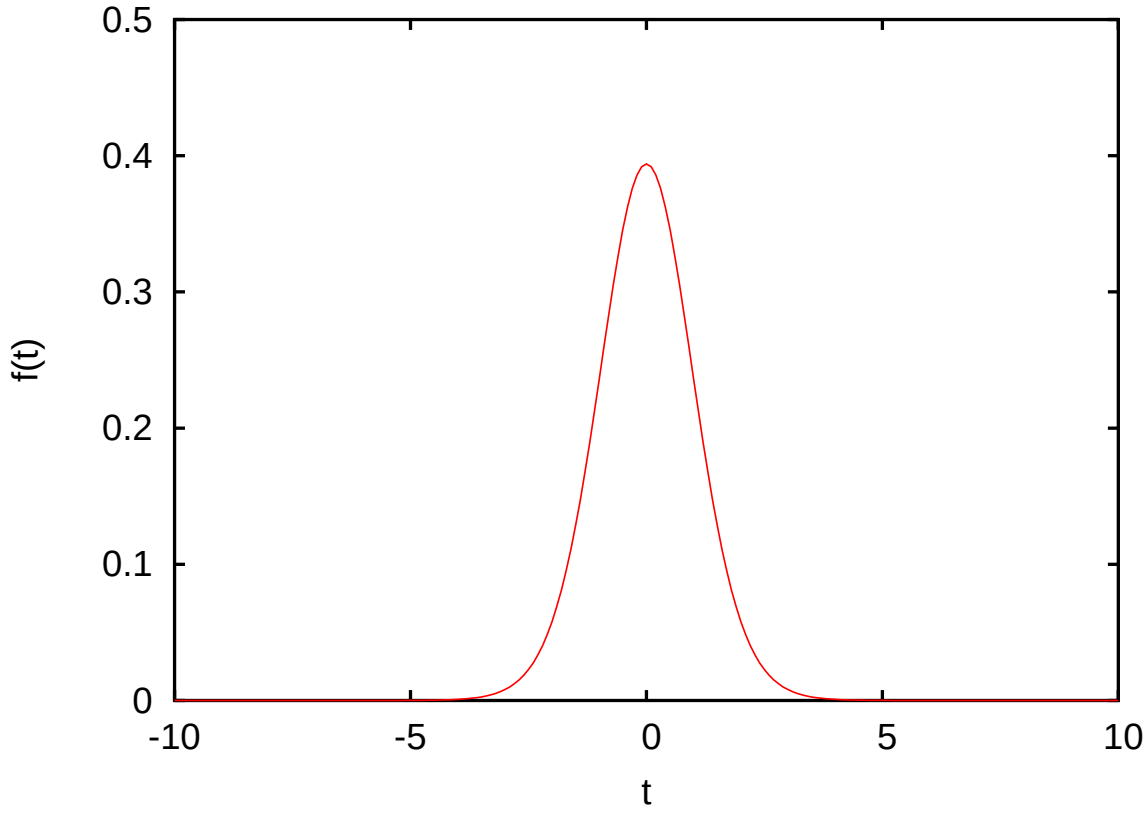
Rysunek: Rozkład prawdopodobieństwa zmiennej $t \sim t(1)$

Rozkład t (Studenta)



Rysunek: Rozkład prawdopodobieństwa zmiennej $t \sim t(5)$

Rozkład t (Studenta)



Rysunek: Rozkład prawdopodobieństwa zmiennej $t \sim t(20)$

Rozkład t (Studenta)

Przykład

Niech zmienne losowe $X_1, \ldots, X_n$ będą niezależne, o takich samych rozkładach $N(\mu, \sigma),\ \sigma > 0$.

Wtedy z twierdzenia Fishera wynika, że:

$$U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1) \qquad \frac{(n-1)S^{*2}}{\sigma^2} \sim \chi^2(n-1)$$

Z definicji rozkładu t Studenta otrzymujemy wtedy dla ilorazu tych zmiennych:

$$t = \frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{1}{n-1}\frac{(n-1)S^{*2}}{\sigma^2}}} = \frac{\overline{X} - \mu}{S^*}\sqrt{n} \sim t(n-1)$$

Rozkład t (Studenta)

Komentarz do przykładu

- właśnie na podstawie tego przekształcenia jest zbudowany test t Studenta – wiadomo, jaki będzie rozkład zmiennej losowej t przy prawdziwości hipotezy $H_0 : \mu = \mu_0$ i nieznanym σ

Rozkład F (Snedecora)

Niech będą dane niezależne zmienne losowe S_x^2, S_y^2 o rozkładach:

$$S_x^2 \sim \chi^2(n) \qquad S_y^2 \sim \chi^2(m)$$

Wtedy o zmiennej losowej:

$$F = \frac{\frac{1}{n} \cdot S_x^2}{\frac{1}{m} \cdot S_y^2}$$

mówimy, że ma **rozkład F Snedecora** o (n,m) stopniach swobody, co oznacza się jako:

$$F \sim F(n, m)$$

Rozkład F (Snedecora)

Komentarz do definicji:

- zmienna o rozkładzie F ma wartości dodatnie
- rozkład F Snedecora powstaje przy dzieleniu nieco przekształconych zmiennych o niezależnych rozkładach χ^2
- liczba stopni swobody rozkładu F pochodzi od liczby stopni swobody obu rozkładów χ^2
- wartości prawdopodobieństw dla rozkładu F są w tablicach

Rozkład F (Snedecora)

Przykład

Niech $X_1, X_2, \ldots, X_n$ i $Y_1, Y_2, \ldots, Y_m$ będą niezależnymi zmiennymi losowymi o rozkładach:

$$X_i \sim N(\mu_x, \sigma) \qquad Y_j \sim N(\mu_y, \sigma)$$

Wtedy o wariancjach z prób wiemy (z twierdzenia Fischera), że:

$$\frac{(n-1)S_x^{*2}}{\sigma^2} \sim \chi^2(n-1)$$

$$\frac{(m-1)S_y^{*2}}{\sigma^2} \sim \chi^2(m-1)$$

i że te zmienne losowe są niezależne.

Rozkład F (Snedecora)

Przykład (c.d.)

W takim razie wstawiając te zmienne do definicji rozkładu F mamy:

$$F = \frac{\frac{(n-1)S_x^{*2}}{(n-1)\sigma^2}}{\frac{(m-1)S_y^{*2}}{(m-1)\sigma^2}} = \frac{S_x^{*2}}{S_y^{*2}} \sim F(n-1, m-1)$$

Rozkład F (Snedecora)

Komentarz do przykładu:

- rozkład F Snedecora w naturalny sposób pojawia się, gdy chcemy testować hipotezę o równości wariancji w dwu niezależnych próbach z dwu rozkładów normalnych
- oczywiście konieczne jest założenie w hipotezie H_0, że wariancje są równe
- średnie w tych rozkładach *nie muszą być ani znane, ani równe*

Dziękuję za uwagę!