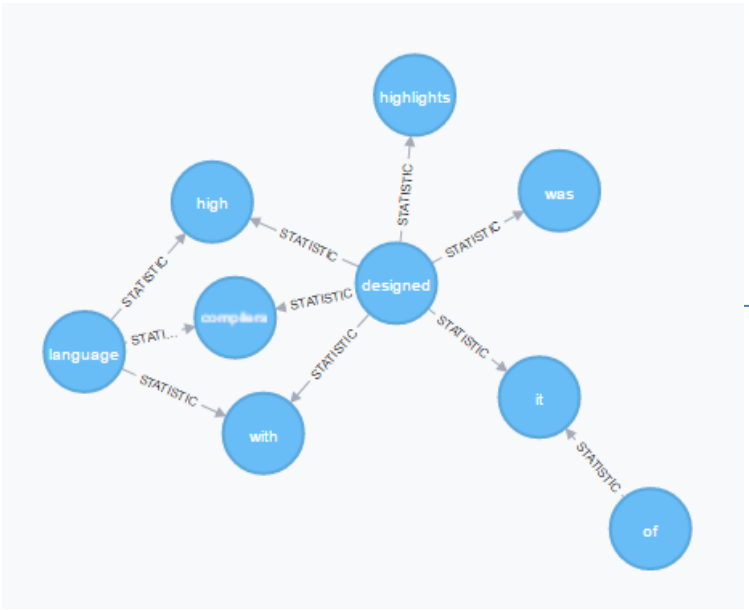
Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

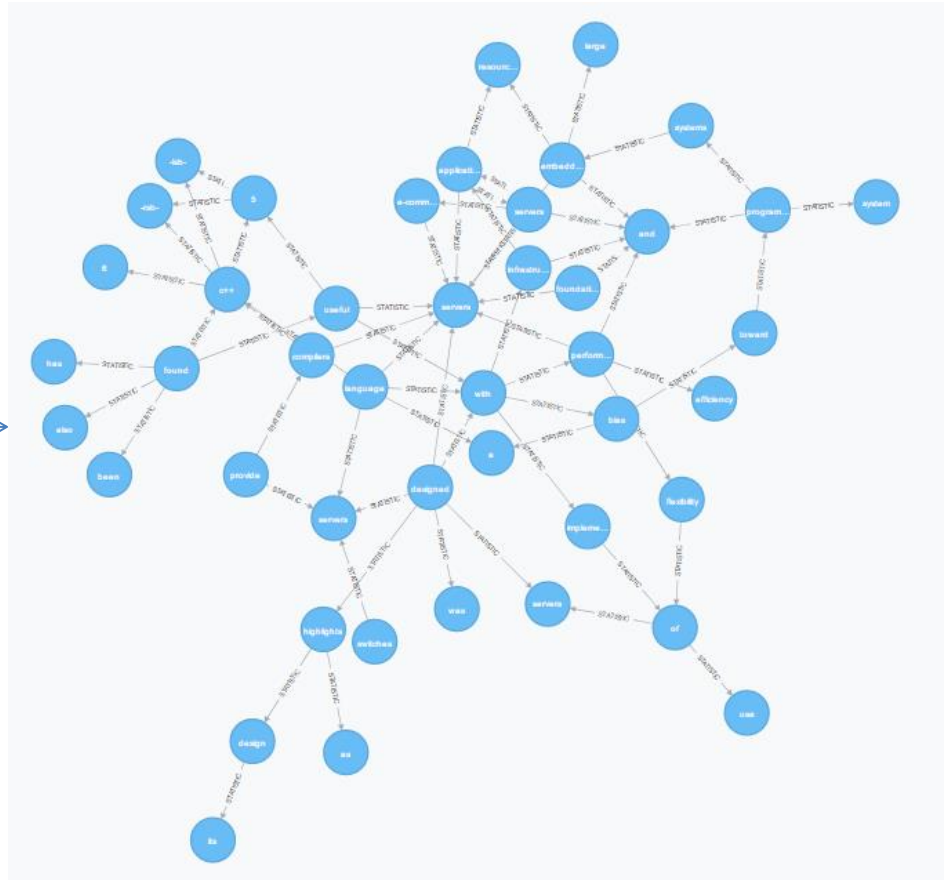
Исследование модели естественного языка, как эволюционирующей сети, с использованием NoSQL системы управления базами данных

Выполнил студент гр. 63601/3 А.С. Крашенинников
Научный руководитель, к.ф.-м.н., доц. А.А. Иванков

Санкт-Петербург
2017

Эволюционная модель





Эволюционная модель(*)

t – параметр “время”(количество лексем)

$k(s,t)$ - степень вершины(количество связей), появившейся в момент s и наблюдаемой в момент t

$c - const$

$$\frac{\partial k(s,t)}{\partial t} = \frac{k(s,t)}{\int_0^t du\, k(u,t)} + 2c\,\frac{k(s,t)\left[\int_0^t du\, k(u,t) - k(s,t)\right]}{\left[\int_0^t du\, k(u,t)\right]^2 - \int_0^t du\, k^2(u,t)} \quad (1)$$

$$k(t,t) = 1 \quad (2)$$

$$\frac{\partial \bar{k}(s,t)}{\partial t} = (1 + 2ct)\,\frac{\bar{k}(s,t)}{\int_0^t du\, \bar{k}(u,t)} \quad (3)$$

$$\bar{k}(t,t) = 1 \quad (4)$$

*- авторы модели Dorogovtsev S.N., Mendes J.F.F.

Эволюционная модель(2)

$$\int_0^t du\, \bar{k}(u,t) = 2t + ct^2 \quad (5)$$

$$K(t) = 2 + ct \quad (6)$$

$$\bar{k}(s,t) = \left(\frac{ct}{cs}\right)^{1/2} \left(\frac{K(t)}{K(s)}\right)^{3/2} \quad (7)$$

$$P_s\left(\bar{k},t\right) = \frac{1}{ct}\frac{K(s)(K(s)-2)}{2K(s)-3}\frac{1}{\bar{k}(s,t)} \quad (8)$$

Модель и оценки(*)

Итоги вычислительного эксперимента:

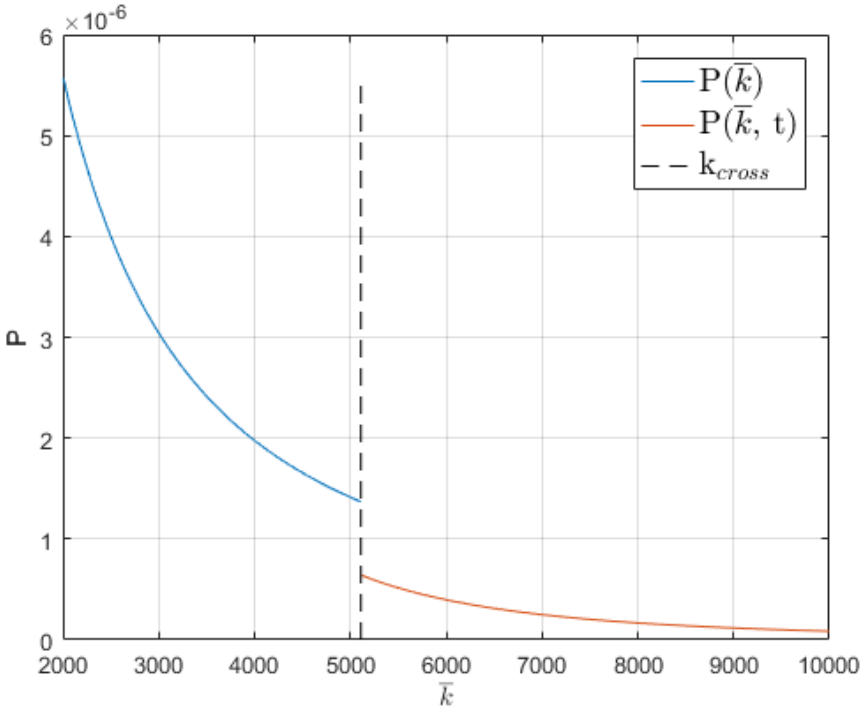
$$|V| = 470000$$

$$K(t) \approx 72 \quad (9)$$

$$k_{cross} \approx \sqrt{ct}(2+ct)^{3/2} \quad (10)$$

$$P(\bar{k}) \simeq \frac{1}{2}\bar{k}^{-3/2}, \bar{k} < k_{cross} \quad (11)$$

$$P(\bar{k},t) \simeq \frac{1}{4}(ct)^3\bar{k}^{-3}, \bar{k} > k_{cross} \quad (12)$$



*- оценки были получены авторами модели

Постановка задачи

- Верификация модели
  - Построение собственной аналитической оценки распределения количества связей
  - Проверка согласия теоретического и эмпирического распределений количества связей
  - Проверка одного из ключевых предположений модели

Эволюционная модель(*)

$$\frac{\partial \bar{k}(s,t)}{\partial t} = (1+2ct)\frac{\bar{k}(s,t)}{\int_0^t du\, \bar{k}(u,t)} \quad (13)$$

$$\bar{k}(t,t) = 1 \quad (14)$$

$$\int_0^t ds\, \bar{k}(s,t) = 2t + ct^2 \quad (15)$$

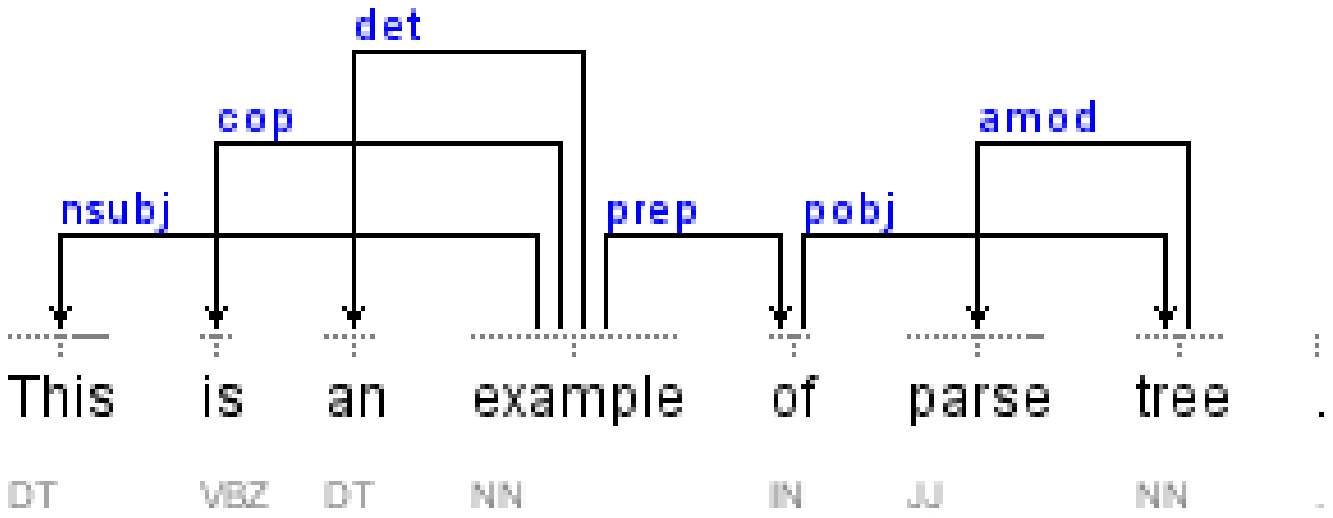
$$\bar{k}(s,t) = t^2\sqrt{\frac{2+ct}{ct}}\, e^C \nrightarrow \left(\frac{ct}{cs}\right)^{1/2}\left(\frac{2+ct}{2+cs}\right)^{3/2} \quad (16)$$

*- выводы проведены в рамках настоящего исследования

Вычислительный эксперимент

Stanford CoreNLP

This is an example of parse tree



Исходные данные(*)

```
1.  <s n="246">
2.      <w c5="AT0" hw="the" pos="ART">The </w>
3.      <w c5="NN1" hw="money" pos="SUBST">money </w>
4.      <w c5="VBD" hw="be" pos="VERB">was </w>
5.      <w c5="NN1" hw="part" pos="SUBST">part </w>
6.      <w c5="PRF" hw="of" pos="PREP">of </w>
7.      <w c5="AT0" hw="the" pos="ART">the </w>
8.      <w c5="NN2" hw="proceed" pos="SUBST">proceeds </w>
9.      <w c5="PRP" hw="from" pos="PREP">from </w>
10.     <w c5="AT0" hw="the" pos="ART">the </w>
11.     <w c5="NN1" hw="sale" pos="SUBST">sale </w>
12.     <c c5="PUN">.</c>
13. </s>
```

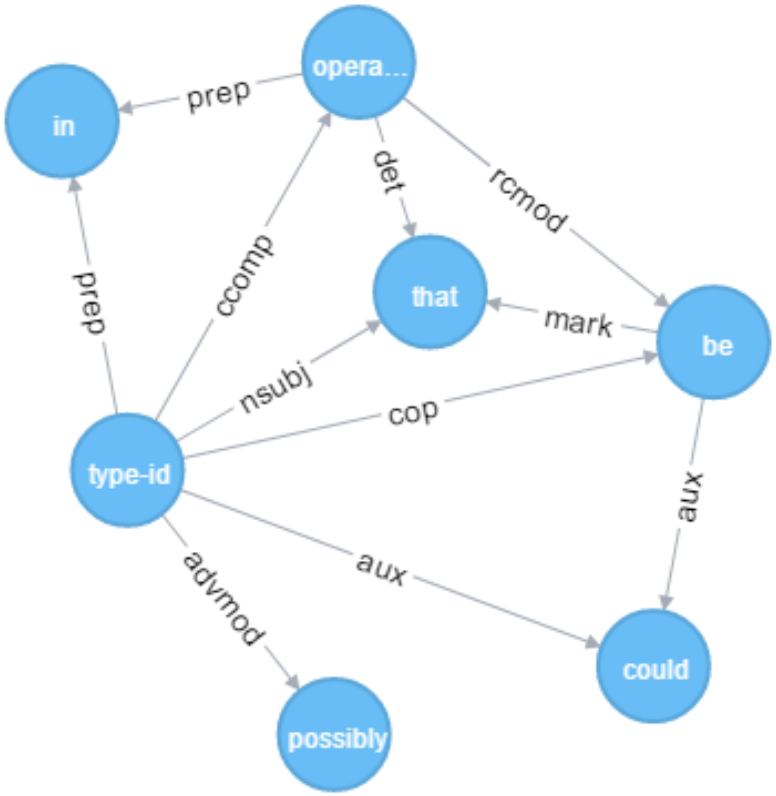
*-исходные данные BNC предоставлены Оксфордским университетом

Графовая СУБД Neo4j

$G(V, E) \stackrel{\text{def}}{=} \langle V, E \rangle$

V – лексемы

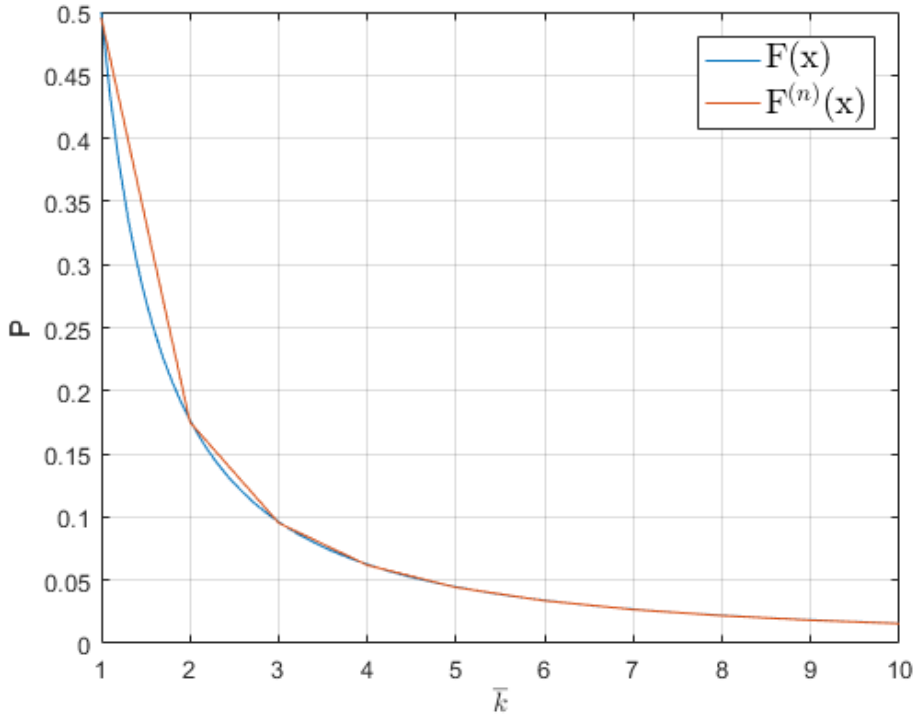
E – семантические связи



Фильтрация семантической сети

1. Знаков препинания и прочих символов, не являющихся символами из алфавита языка (‘<’,’>’,’&’ и т.д.)
2. Символов, не принадлежащих алфавиту английского языка [a-z, A-Z]
3. Артиклей (‘a’,’the’ и т.д.)
4. Предлогов

Проверка статистической гипотезы о согласии



$F(x)$ – теоретическая функция распределения
$F^{(n)}(x)$ - выборочная функция распределения

Критерий Колмогорова-Смирнова

$$P\left\{F(x) - d_\alpha \le F^{(n)}(x) \le F(x) + d_\alpha\right\} = 1 - \alpha \qquad (17)$$

где:

- $F^{(n)}(x)$ - выборочная функция распределения,
- $F(x)$ – теоретическая функция распределения,
- α - уровень значимости,
- d_α - критическое значение

$$d_\alpha \approx \sqrt{-\frac{1}{2} ln \frac{\alpha}{2}} / \sqrt{n}$$

оценка:

$$\hat{P} = \frac{n_\text{П}}{n} \qquad (18)$$

где:

- $n_\text{П}$ - количество попаданий ординат выборочной функции в доверительный интервал

Результаты проверки статистической гипотезы о согласии распределений(*)

Шаги этапа фильтрации	$\hat{P}$
Без фильтрации	0,954
Знаки препинания и не алфавитные символы (‘<’,’>’,’&’ и т.д.)	0,958
Символы не английского алфавита	0,962
Артикли	0,962
Предлоги	0,965

*- таблица составлена на основании проведенного вычислительного эксперимента. Здесь $\hat{P}$ - соотв. (18)

Проверка предположения модели

$$\frac{\partial \bar{k}(s,t)}{\partial t} = (1 + 2ct)\frac{\bar{k}(s,t)}{\int_0^t du\, \bar{k}(u,t)} \quad (19)$$

$$c = const \quad (20)$$

Корпус	Количество узлов	Количество связей	Среднее количество связей
C++98	5123	66953	26.00
C++11	5235	72392	27.51
Σ	6146	90622	29.33

Проверка предположения модели

$$\frac{\partial k_0}{\partial t} = (1 + 2\hat{c}t_2)\frac{k(t_0, t_2)}{2t_2 + \hat{c}{t_2}^2}$$
$$\frac{\partial k_1}{\partial t} = (1 + 2\widehat{c'}t_2)\frac{k(t_1, t_2)}{2t_2 + \widehat{c'}{t_2}^2}$$

где:

- t_0 - соотв. сети, построенной на основе C++98
- t_1 - соотв. сети, построенной на основе C++11
- t_2 - соотв. общей сети

Оценка:

$$\hat{c} = 0.37$$
$$\widehat{c'} = 0.13$$

Итоги исследования

- Проведена верификация одной эволюционной модели ЕЯ
  - Проведены аналитические выводы для того, чтобы воспроизвести теоретические оценки авторов
  - Воспроизведены эмпирические оценки авторов модели на корпусе BNC
  - Проведена проверка одного из ключевых предположений модели
- Разработано ПО(Java) для решения подобных задач(NLP)

Спасибо за внимание!

-

- T

```
<sentence id="1">
  <tokens>
    <token id="1">
      <word>Road</word>
      <POS>NNP</POS>
    </token>
    <token id="2">
      <word>Street</word>
      <POS>NNP</POS>
    </token>
    …
  </tokens>
  <dependencies type="basic-dependencies">
    <dep type="root">
      <governor idx="0">ROOT</governor>
      <dependent idx="7">Road</dependent>
    </dep>
    <dep type="conj">
      <governor idx="7">Road</governor>
      <dependent idx="10">Street</dependent>
    </dep>
    …
  </dependencies>
</sentence>
```

```
<s n="246">
  <w c5="AT0" hw="the" pos="ART">The </w>
  <w c5="NN1" hw="money" pos="SUBST">money </w>
  <w c5="VBD" hw="be" pos="VERB">was </w>
  <w c5="NN1" hw="part" pos="SUBST">part </w>
  <w c5="PRF" hw="of" pos="PREP">of </w>
  <w c5="AT0" hw="the" pos="ART">the </w>
  <w c5="NN2" hw="proceed" pos="SUBST">proceeds </w>
  <w c5="PRP" hw="from" pos="PREP">from </w>
  <w c5="AT0" hw="the" pos="ART">the </w>
  <w c5="NN1" hw="sale" pos="SUBST">sale </w>
   <c c5="PUN">.</c>
</s>
```

$$\frac{\partial \bar{k}(s,t)}{\partial t} = (1 + 2ct)\frac{\bar{k}(s,t)}{\int_0^t du\, \bar{k}(u,t)} \qquad (21)$$

Применяем $\int_0^t ds$ к левой и правой частям:

$$(21) \rightarrow \int_0^t ds \frac{\partial \bar{k}(s,t)}{\partial t}$$

Используем ф.Лейбница:

$$\int_0^t ds \frac{\partial \bar{k}(s,t)}{\partial t} = \frac{\partial}{\partial t}\int_0^t ds\, \bar{k}(s,t) - \bar{k}(t(t),t)\frac{\partial t}{\partial t} + \bar{k}(0(t),t)\frac{\partial 0}{\partial t}$$
$$= \frac{\partial}{\partial t}\int_0^t ds\, \bar{k}(s,t) - \bar{k}(t,t) \qquad (22)$$

$$(21) \rightarrow \int_0^t (1+2ct)\frac{\bar{k}(s,t)}{\int_0^t du\ \bar{k}(u,t)}ds =$$

$$(1+2ct)\frac{\int_0^t ds\ \bar{k}(s,t)}{\int_0^t du\ \bar{k}(u,t)} = 1+2ct \qquad (23)$$

$$(22),(23) \rightarrow \frac{\partial}{\partial t}\int_0^t ds\ \bar{k}(s,t) - \bar{k}(t,t) = 1+2ct$$

$$\Longrightarrow \frac{\partial}{\partial t}\int_0^t ds\ \bar{k}(s,t) = 2+2ct \Longrightarrow \int_0^t ds\ \bar{k}(s,t) = 2t+ct^2$$

$$\frac{\partial k_0}{\partial t} = (1 + 2ct_2)\frac{k(t_0, t_2)}{2t_2 + ct_2{}^2} \Longrightarrow$$

$$\frac{\partial k_0}{\partial t} = (1 + 2ct_2)\frac{\left(\frac{2 + ct_2}{2 + ct_0}\right)}{t_2(2 + ct_2)}\sqrt{\frac{t_0}{t_2}}\sqrt{\frac{2 + ct_2}{2 + ct_0}} \Longrightarrow$$

$$\frac{\partial k_0}{\partial t}\frac{t_2(2 + ct_0)}{1 + 2ct_2} = \sqrt{\frac{t_0}{t_2}}\sqrt{\frac{2 + ct_2}{2 + ct_0}} \Rightarrow$$

$$\left(\frac{\partial k_0}{\partial t}\right)^2 \frac{{t_2}^2(4+4ct_0+c^2{t_0}^2)}{1+4ct_2+4c^2{t_2}^2} = \frac{t_0}{t_2}\frac{2+ct_2}{2+ct_0} \Longrightarrow$$

$$\left(\left(\frac{\partial k_0}{\partial t}\right)^2 {t_0}^3{t_2}^3 - 4t_0{t_2}^3\right)c^3 + \left(6\left(\frac{\partial k_0}{\partial t}\right)^2 {t_0}^2{t_2}^3 - 12t_0{t_2}^2\right)c^2$$

$$+\left(12\left(\frac{\partial k_0}{\partial t}\right)^2 t_0{t_2}^3 - 9t_0t_2\right)c + 8\left(\frac{\partial k_0}{\partial t}\right)^2 {t_2}^3 - 2t_0 = 0$$

Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

Диссертация на соискание степени магистра

Тема: Исследование одной модели естественного языка, как эволюционирующей сети, средствами NoSQL системы управления базами данных

Выполнил студент гр. 63601/3 А.С. Крашенинников
Научный руководитель, к.ф.-м.н., доц. А.А. Иванков

Санкт-Петербург
2017

Критерий Колмогорова-Смирнова(д.)

$$D_n^+ = \max_{1 \le i \le n} \left(\frac{i}{n} - t_i \right)$$

$$D_n^- = \max_{1 \le i \le n} \left(t_i - \frac{i-1}{n} \right)$$

$$D_n = \max_{1 \le i \le n} (D_n^+, D_n^-)$$

$$t_i = F_0(x_{(i)})$$

Где:

- t_i - значение гипотетической функции распределения, взятой в точке i вариационного ряда.