Анализ данных для лингвистов

Г. А. Мороз

Оглавление

Глава 1

О курсе

Материалы для курса Анализа данных для лингвистов, Школа лингвистики НИУ ВШЭ.

```
library(tidyverse)
```

Глава 2

Распределения

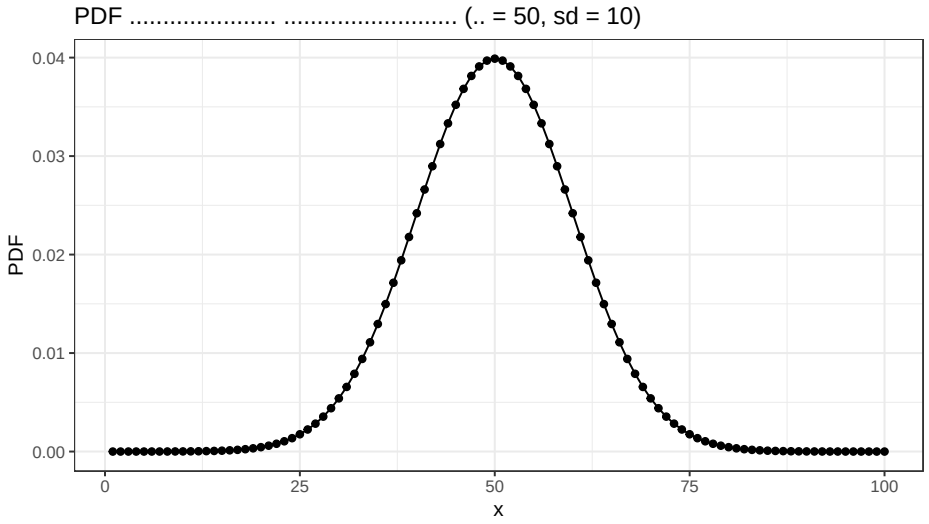
2.1 Распределения в R

В R встроено какое-то количество известных распределений. Все они представлены четырьмя функциями:

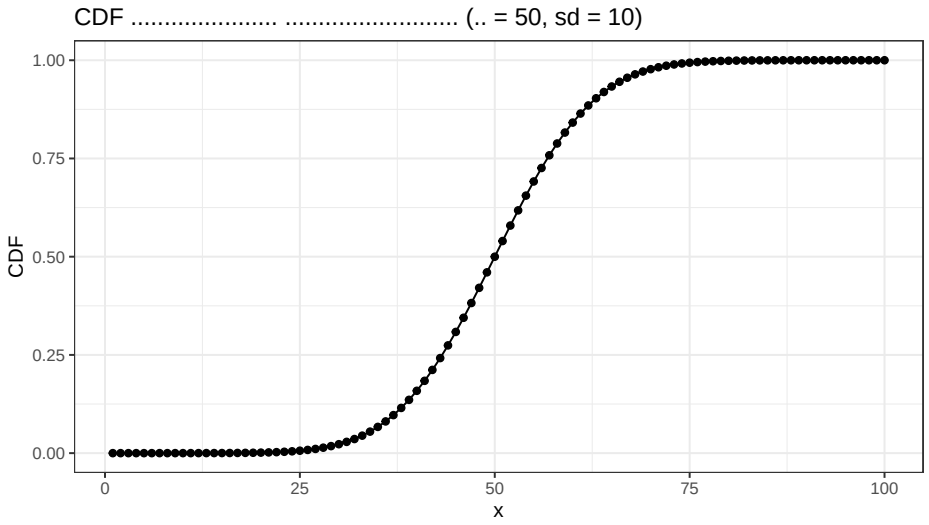
- `d...` (функция плотности, probability density function),
- `p...` (функция распределения, cumulative distribution function) — интеграл площади под кривой от начала до указанной квантили
- `q...` (обратная функции распределения, inverse cumulative distribution function) — значение p-той квантили распределения
- и `r...` (рандомные числа из заданного распределения).

Рассмотрим все это на примере нормального распределения.

```
tibble(x = 1:100,
       PDF = dnorm(x = x, mean = 50, sd = 10)) %>%
  ggplot(aes(x, PDF))+
  geom_point()+
  geom_line()+
  labs(title = "PDF                ( = 50, sd = 10)")
```

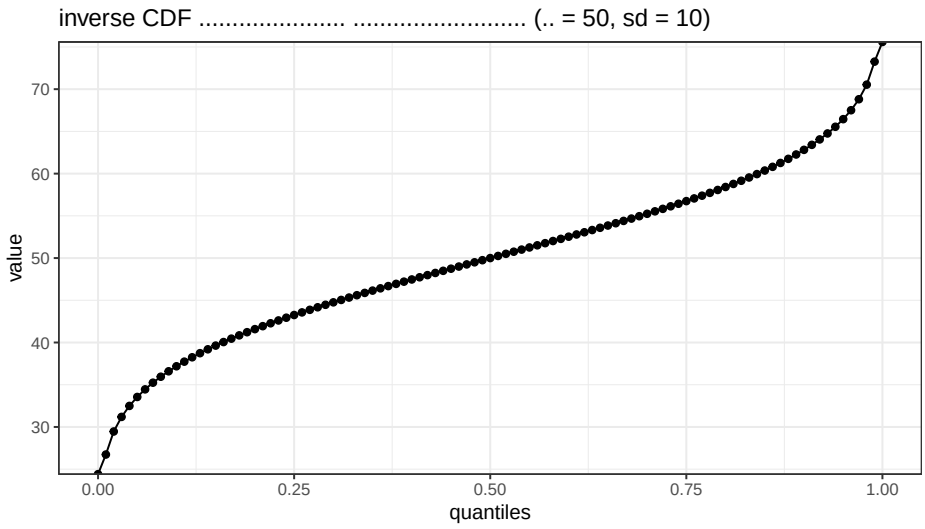



```
tibble(x = 1:100,
       CDF = pnorm(x, mean = 50, sd = 10)) %>%
  ggplot(aes(x, CDF))+
  geom_point()+
  geom_line()+
  labs(title = "CDF                ( = 50, sd = 10)")
```

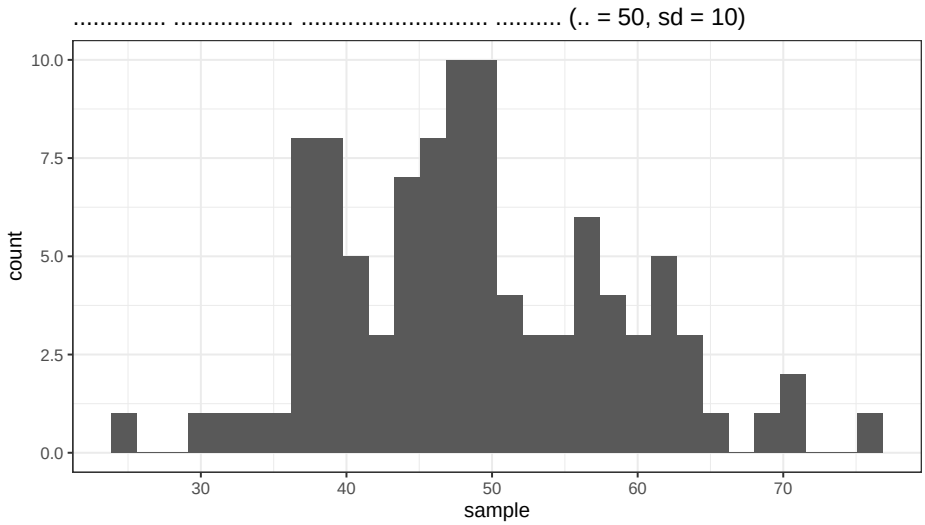



```
tibble(quantiles = seq(0, 1, by = 0.01),
       value = qnorm(quantiles, mean = 50, sd = 10)) %>%
  ggplot(aes(quantiles, value))+
```

```
  geom_point()+
  geom_line()+
  labs(title = "inverse CDF                ( = 50, sd = 10)")
```




```
tibble(sample = rnorm(100, mean = 50, sd = 10)) %>%
  ggplot(aes(sample))+
  geom_histogram()+
  labs(title = "                          ( = 50, sd = 10)")
```



Если не использовать `set.seed()`, то результат работы рандомизатора нельзя будет

повторить.

Какое значение имеет 25% квантиль нормального распределения со средним в 20 и стандартным отклонением 90? Ответ округлите до трех знаков после запятой.

Данные из базы данных фонетических инвентарей PHOIBLE [@phoible], достаточно сильно упрощая, можно описать нормальным распределением со средним 35 фонем и стандартным отклонением 13. Если мы ничего не знаем про язык, оцените с какой вероятностью, согласно этой модели произвольно взятый язык окажется в промежутке между 50 фонемами и 25? Ответ округлите до трех знаков после запятой.

Какие есть недостатки у модели из предыдущего задания?

2.2 Категориальные переменные

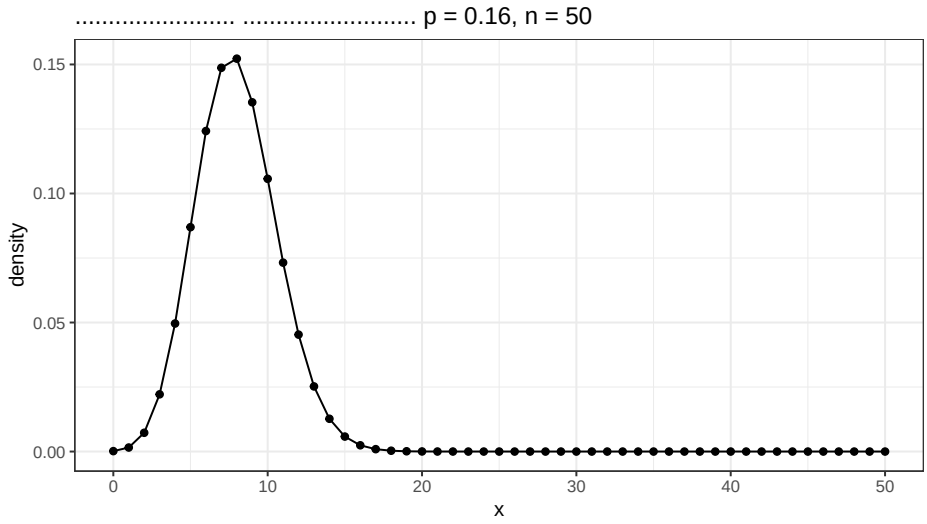
2.2.1 Биномиальное распределение

Биномиальное распределение — распределение количетсва успехов эксперементов Бернулли из n попыток с вероятностью успеха p.

$$P(k|n,p) = \frac{n!}{k!(n-k)!} \times p^k \times (1-p)^{n-k} = \binom{n}{k} \times p^k \times (1-p)^{n-k}$$

$$0 \leq p \leq 1; n, k > 0$$

```
tibble(x = 0:50,
       density = dbinom(x = x, size = 50, prob = 0.16)) %>%
  ggplot(aes(x, density))+
  geom_point()+
  geom_line()+
  labs(title = "                p = 0.16, n = 50")
```



Немного упрощая данные из статьи [@rosenbach03], можно сказать...

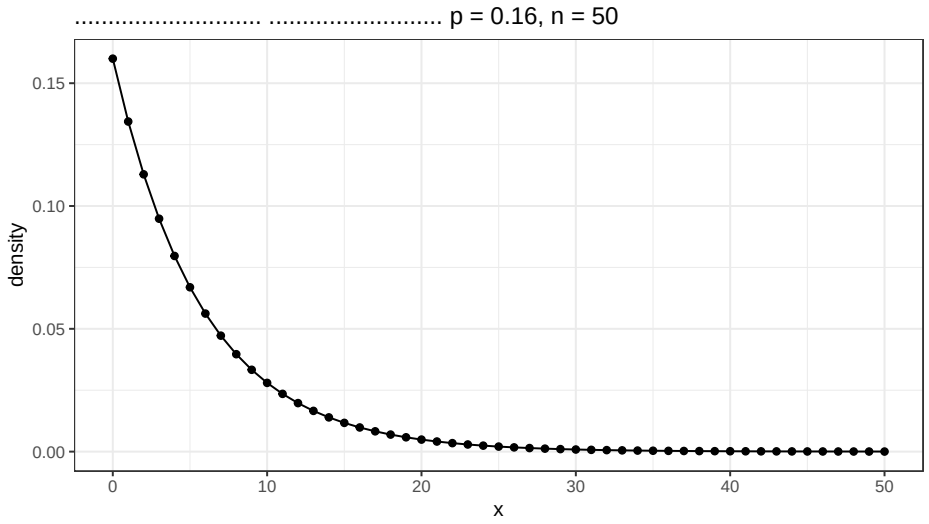
2.2.2 Геометрическое распределение

Геометрическое распределение — распределение количетсва эксперементов Бернулли с вероятностью успеха p до первого успеха.

$$P(k|p) = (1-p)^k \times p$$

$$k \in \{1, 2, \dots\}$$

```
tibble(x = 0:50,
       density = dgeom(x = x, prob = 0.16)) %>%
  ggplot(aes(x, density))+
  geom_point()+
  geom_line()+
  labs(title = "                p = 0.16, n = 50")
```



2.2.3 Распределение Пуассона

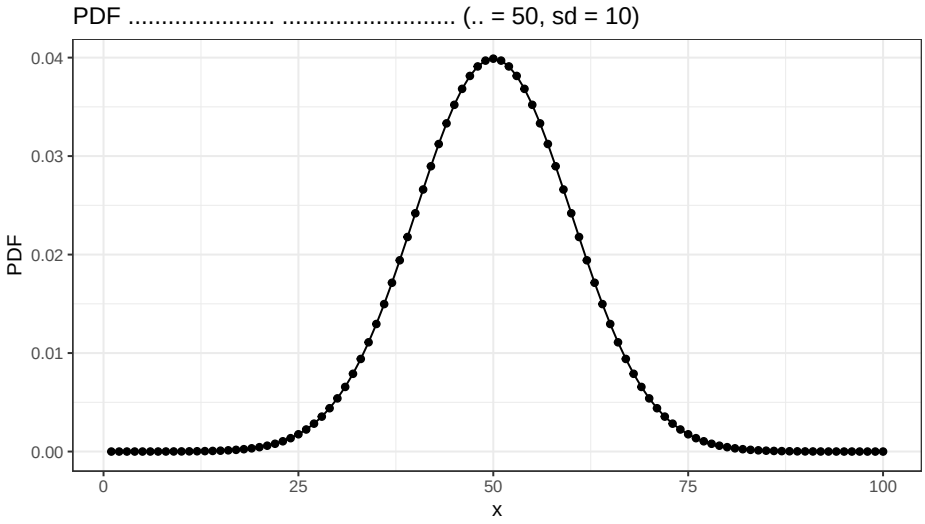
2.3 Числовые переменные

2.3.1 Нормальное распределение

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu \in \mathbb{R}; \sigma^2 > 0$$

```
tibble(x = 1:100,
       PDF = dnorm(x = x, mean = 50, sd = 10)) %>%
  ggplot(aes(x, PDF))+
  geom_point()+
  geom_line()+
  labs(title = "PDF                ( = 50, sd = 10)")
```

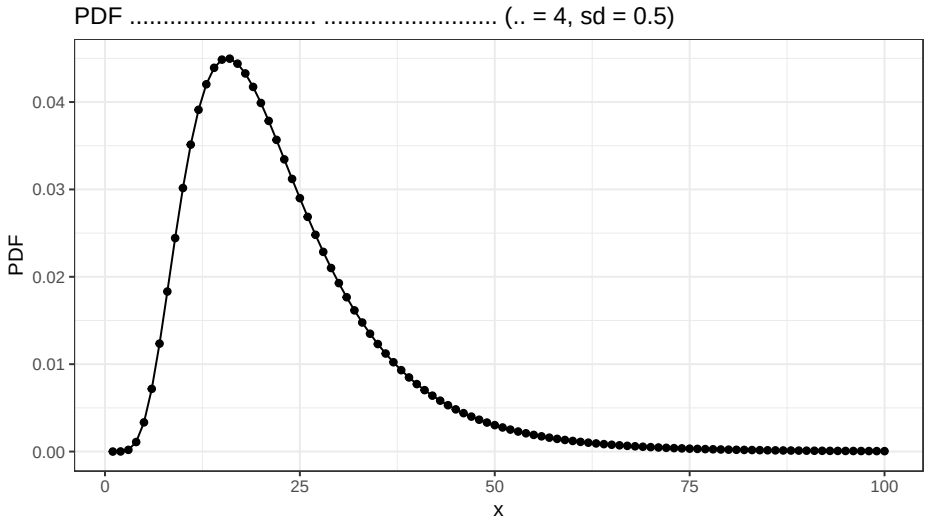


2.3.2 Логнормальное распределение

$$P(x) = \frac{1}{\sqrt{x\sigma 2\pi}} \times e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

$$\mu \in \mathbb{R}; \sigma^2 > 0$$

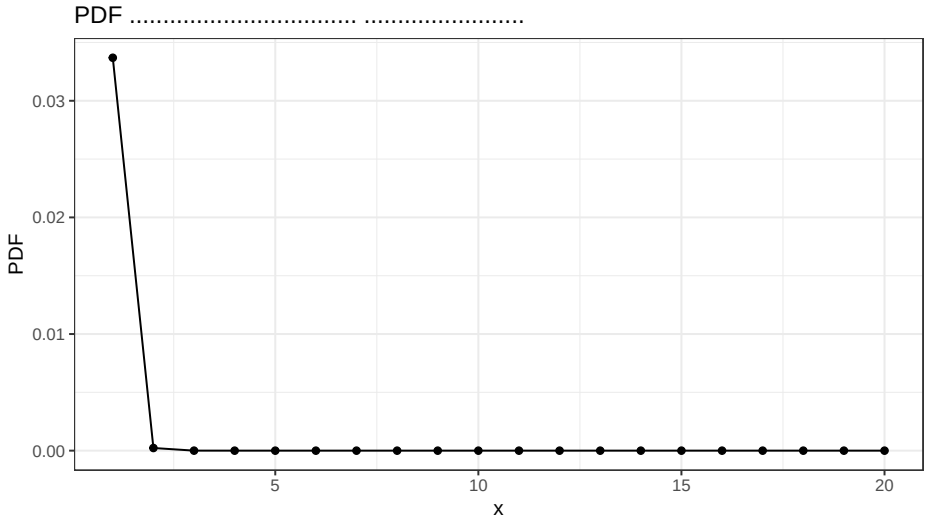
```
tibble(x = 1:100,
       PDF = dlnorm(x = x, mean = 3, sd = 0.5)) %>%
  ggplot(aes(x, PDF))+
  geom_point()+
  geom_line()+
  labs(title = "PDF                  (  = 4, sd = 0.5)")
```



2.3.3 Экспоненциальное распределение

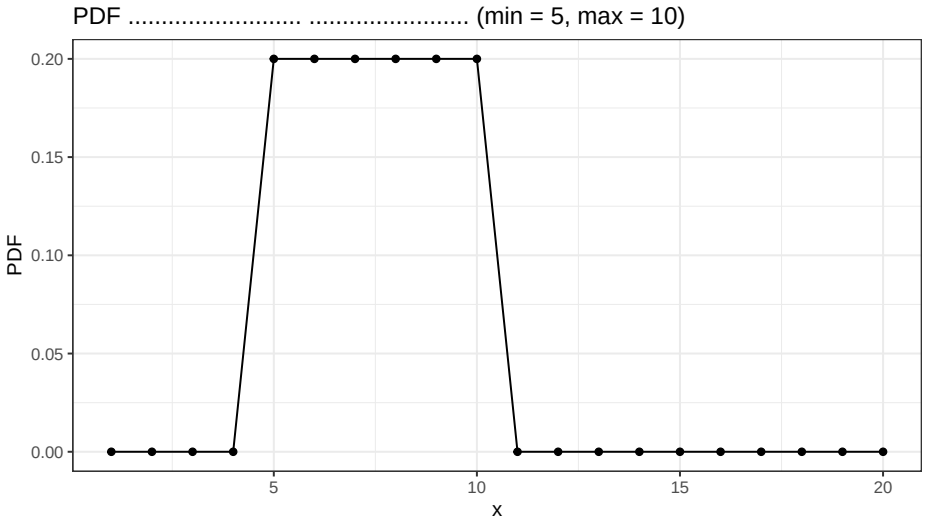
$$P(x) = \lambda \times e^{-\lambda x}$$

```
tibble(x = 1:20,
       PDF = dexp(x = x, rate = 5)) %>%
  ggplot(aes(x, PDF))+
  geom_point()+
  geom_line()+
  labs(title = "PDF                    ")
```



2.3.4 Унимодальное распределение

```
tibble(x = 1:20,
       PDF = dunif(x = x, min = 5, max = 10)) %>%
  ggplot(aes(x, PDF))+
  geom_point()+
  geom_line()+
  labs(title = "PDF                  (min = 5, max = 10)")
```



2.3.5 Что еще почитать про распределения?

В интернете много ресурсов, но вот еще есть вот этот[1]. А здесь[2] можно найти соответствия распределений и сопряжённым к ним априорных распределений.

[1] http://www.math.wm.edu/~leemis/chart/UDR/UDR.html
[2] https://en.wikipedia.org/wiki/Conjugate_prior