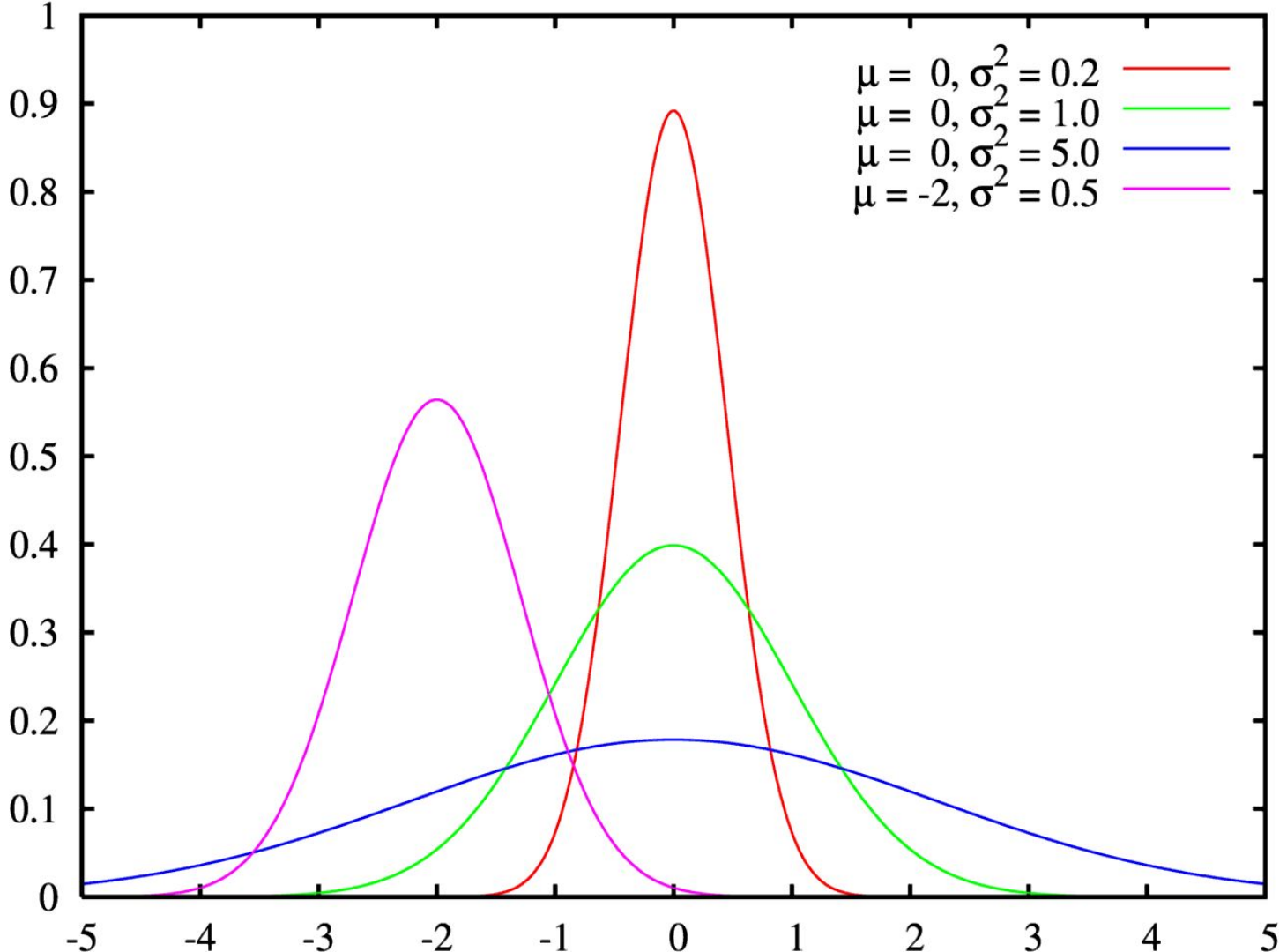
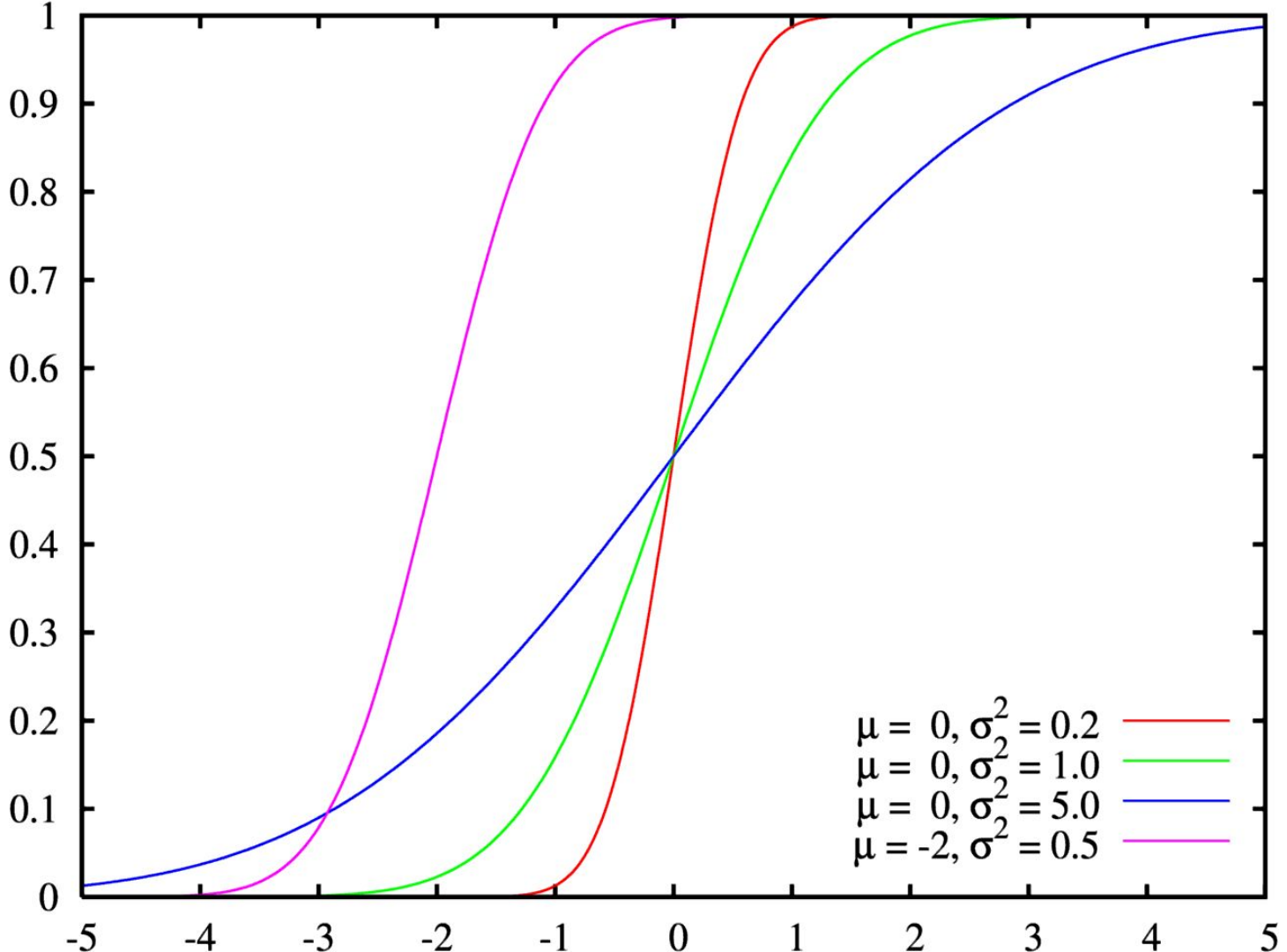
Введение в статистику

Распределение вероятностей — это закон, описывающий область значений случайной величины и вероятности их исхода (появления).

Функция и плотность распределения


μ = 0, σ² = 0.2
μ = 0, σ² = 1.0
μ = 0, σ² = 5.0
μ = -2, σ² = 0.5
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
μ = 0, σ² = 0.2
μ = 0, σ² = 1.0
μ = 0, σ² = 5.0
μ = -2, σ² = 0.5

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

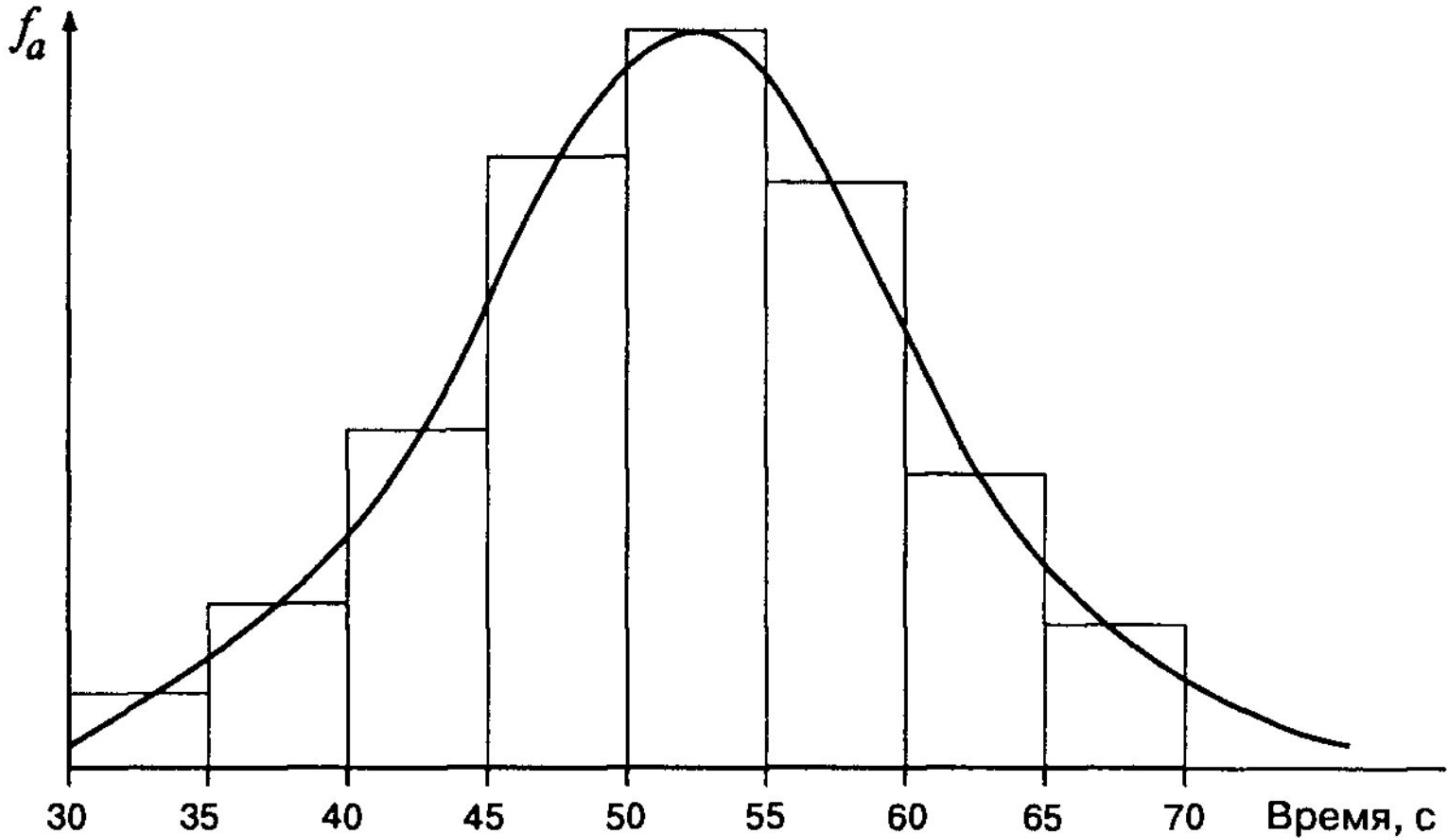


Рис. 3.4. Гистограмма и сглаженный график распределения частот времени решения тестовой задачи (по данным табл. 3.3)

Центра́льные преде́льные теоре́мы (ЦПТ) — класс теорем в теории вероятностей, утверждающих, что сумма достаточно большого количества слабо зависимых случайных величин, имеющих примерно одинаковые масштабы (ни одно из слагаемых не доминирует, не вносит в сумму определяющего вклада), имеет распределение, близкое к нормальному.

Меры центральной тенденции

Мода

Медиана

Среднее

Меры изменчивости

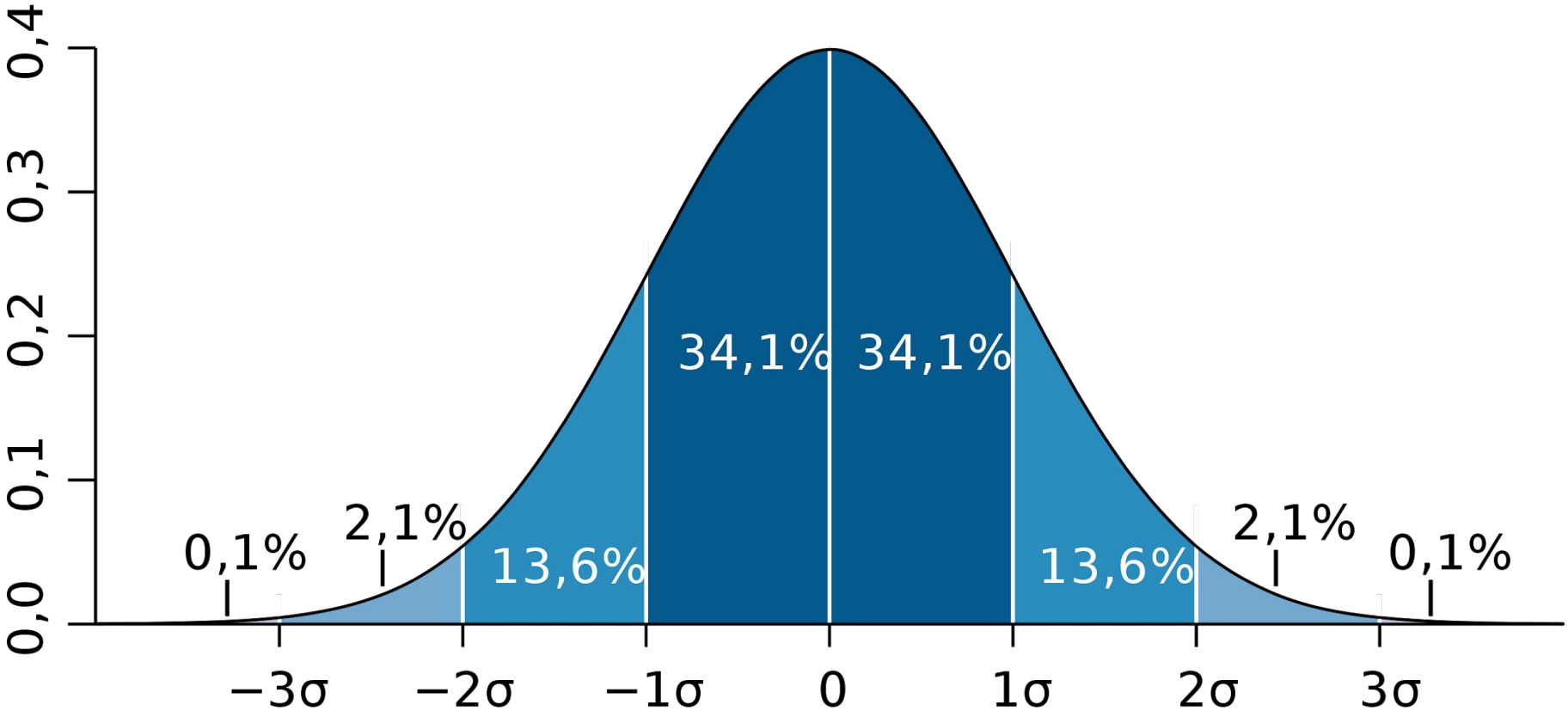
Размах

$$R = x_{\max} - x_{\min}.$$

$$P_{90} - P_{10}$$

Стандартное отклонение

$$\sigma_x = \sqrt{D_x} = \sqrt{\frac{\sum_i (x_i - M_x)^2}{N-1}}.$$


0,4
0,3
0,2
0,1
0,0
0,1%
2,1%
13,6%
34,1%
34,1%
13,6%
2,1%
0,1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

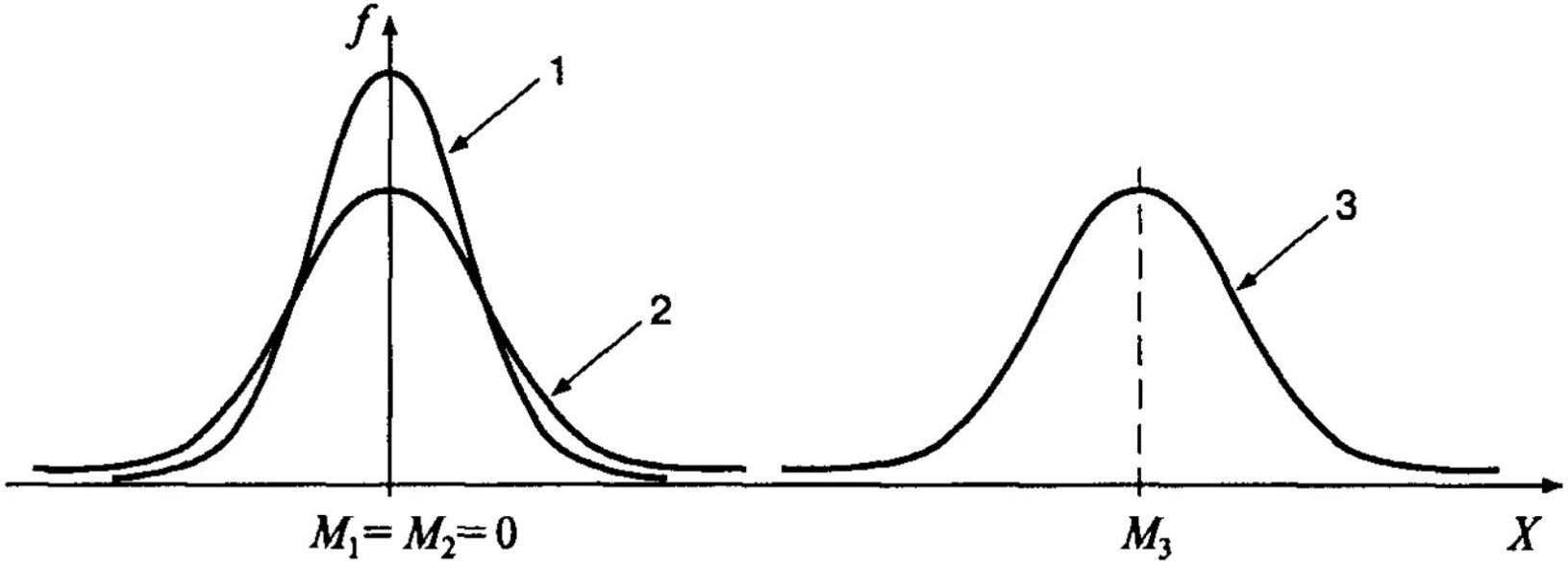
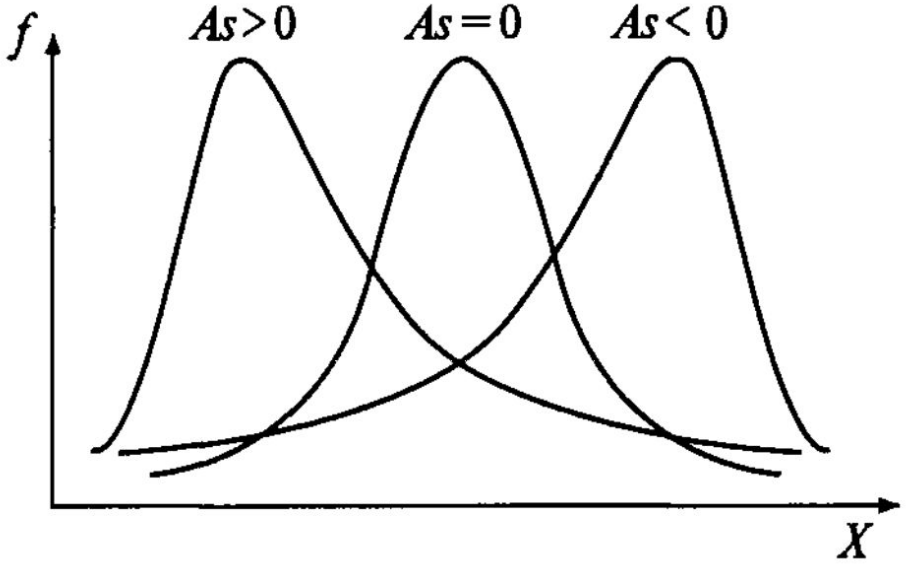


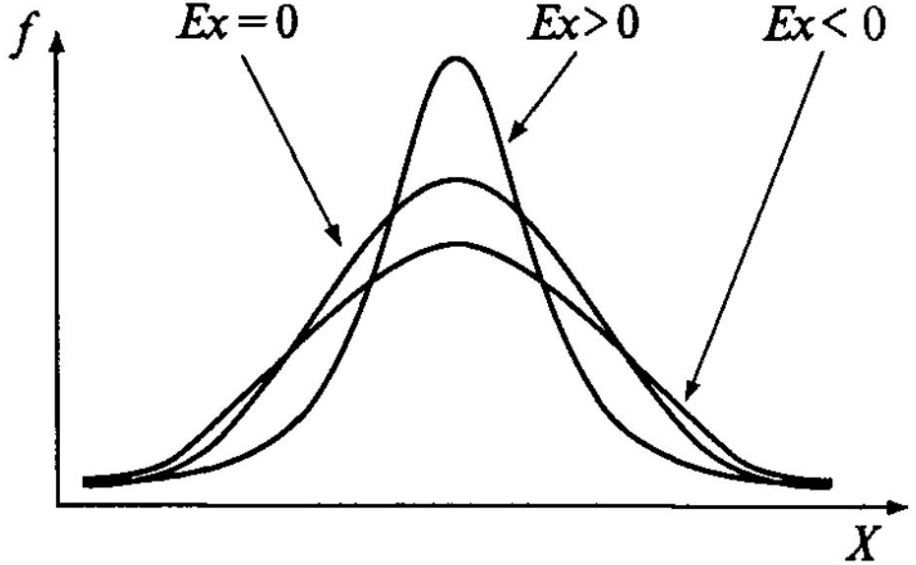
Рис. 4.1. Графики распределения частот: с разной дисперсией ($D_1 < D_2$), одинаковой дисперсией ($D_2 = D_3$) и разными средними арифметическими ($M_2 < M_3$)

Z - преобразование

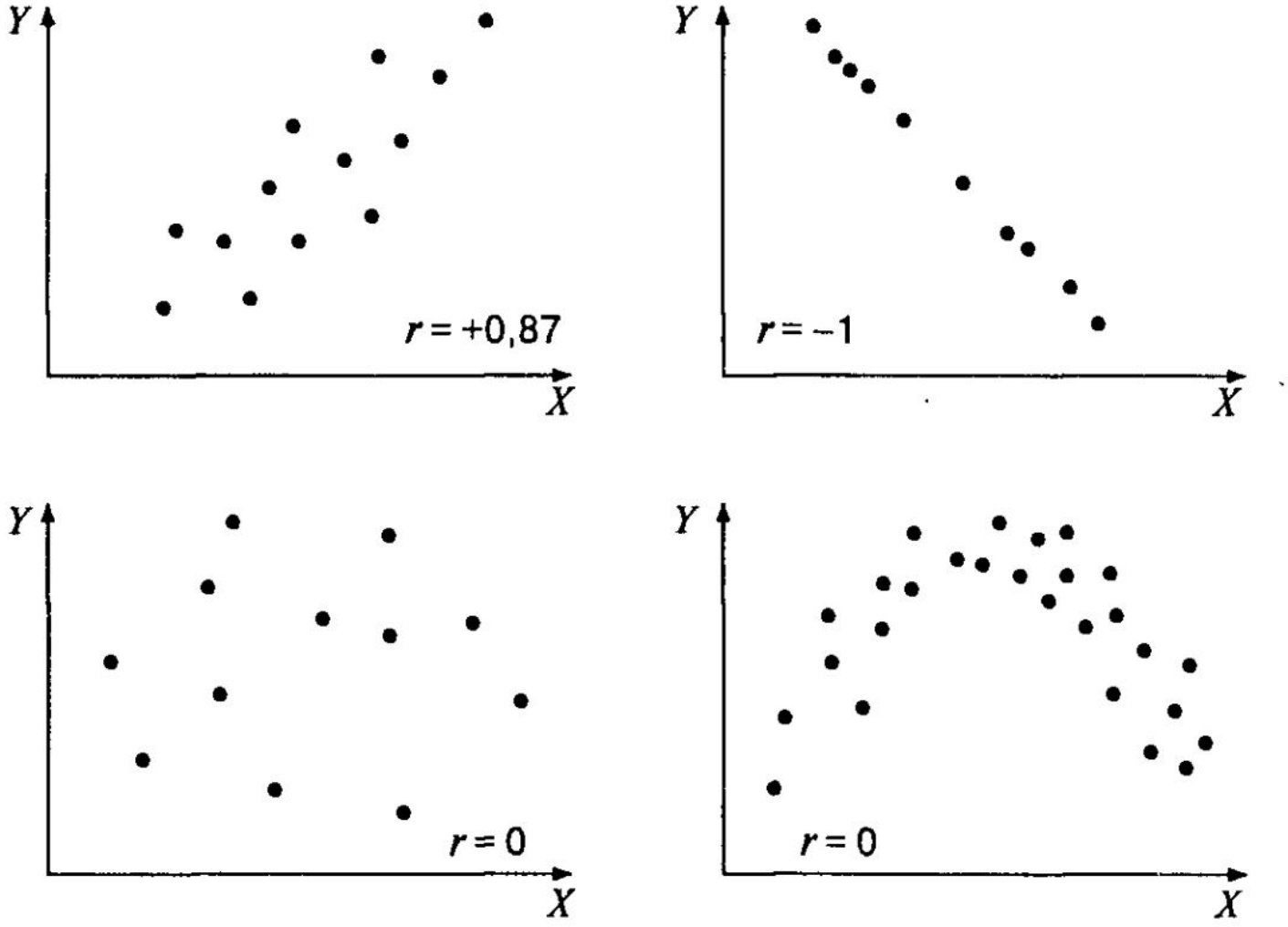
$$z_i = \frac{x_i - M_x}{\sigma_x}.$$

Асимметрия и эксцесс





Корреляция


Y
X
r = +0,87
Y
X
r = –1
Y
X
r = 0
Y
X
r = 0

r-Пирсона

$$r_{xy} = \frac{\sum_{i=1}^{N} (x_i - M_x)(y_i - M_y)}{(N-1)\sigma_x \sigma_y}$$

r-Спирмена

$$r_s = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)},$$

```scala
import org.apache.spark.ml.linalg.{Matrix, Vectors}
import org.apache.spark.ml.stat.Correlation
import org.apache.spark.sql.Row

val data = Seq(
  Vectors.sparse(4, Seq((0, 1.0), (3, -2.0))),
  Vectors.dense(4.0, 5.0, 0.0, 3.0),
  Vectors.dense(6.0, 7.0, 0.0, 8.0),
  Vectors.sparse(4, Seq((0, 9.0), (3, 1.0)))
)

val df = data.map(Tuple1.apply).toDF("features")
val Row(coeff1: Matrix) = Correlation.corr(df, "features").head
println(s"Pearson correlation matrix:\n $coeff1")

val Row(coeff2: Matrix) = Correlation.corr(df, "features", "spearman").head
println(s"Spearman correlation matrix:\n $coeff2")
```

Проверка гипотез

H0 - основная гипотеза

H1 - альтернативная гипотеза

t-Стьюдента для одной выборки

H_0: $\bar{M}_x = A$.

$$t_э = \frac{|M - A|}{\sigma / \sqrt{N}},$$

P-Value Approach

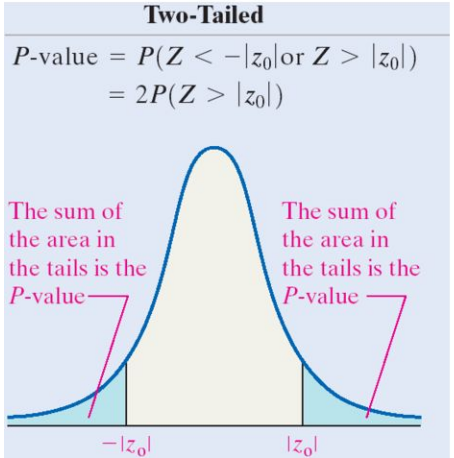
Assume that the null hypothesis is true.

The P-Value is the probability of observing a sample mean that is as or more extreme than the observed.

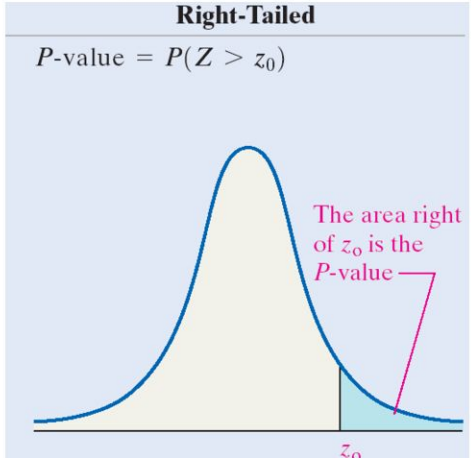
How to compute the P-Value for each type of test:

Step 1: Compute the test statistic $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

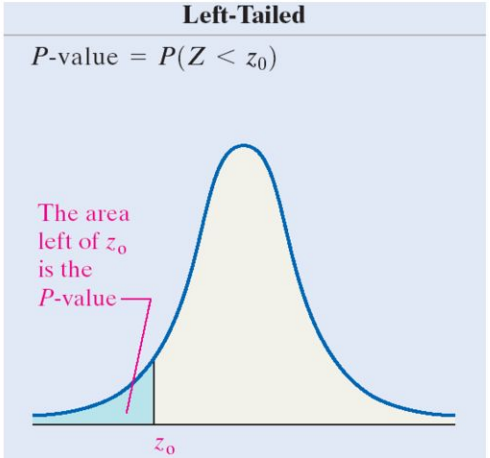
Two-tail



Right Tail



Left Tail



t-Стьюдента для двух независимых выборок

$$H_0: \bar{M}_1 = \bar{M}_2.$$

$$t_э = \frac{|M_1 - M_2|}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$