

**Задача 1.** Дана выборка: 4, 18, 9, 10, -1, 9, 25, 18, 9

- Запишите вариационный ряд и ранги наблюдений.
- Найдите медиану этой выборки. Проинтерпретируйте полученное значение.
- Найдите верхний и нижний квартили этой выборки.
- Проверьте, есть ли в выборке нетипичные значения (выбросы). Если нет, напишите, что их нет, если есть – перечислите.
- Постройте для данной выборки «ящик с усами».
- Постройте гистограмму для данной выборки, выбрав стартовым значением минимальное значение в выборке и интервал группировки равный 4.
- Найдите выборочное среднее и выборочное стандартное отклонение (несмещенную оценку).

**Решение.**

a. Вариационный ряд: -1, 4, 9, 9, 9, 10, 18, 18, 25

Ранги:  $R_1 = 2, R_2 = 7.5, R_3 = 4, R_4 = 6, R_5 = 1, R_6 = 4, R_7 = 9, R_8 = 7.5, R_9 = 4$ .

b. Медиана:  $med(X) = 9$ ; Верхний квартиль:  $Q_3 = 18$ ; Нижний квартиль  $Q_1 = 9$ .

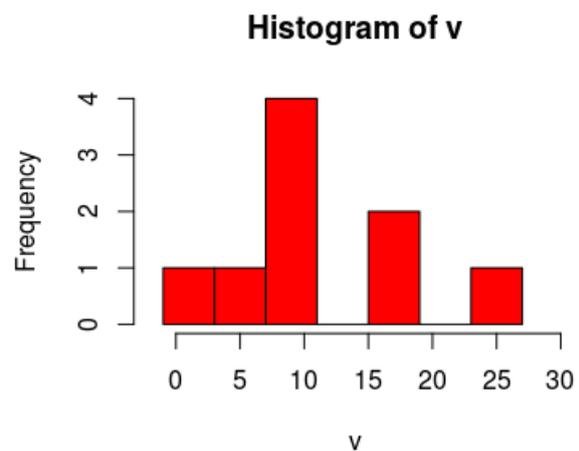
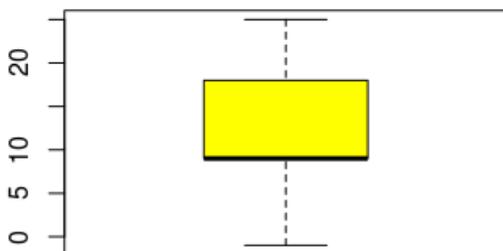
c. Границы типичных значений:

$$[9 - 1.5 \cdot (18 - 9); 18 + 1.5 \cdot (18 - 9)]$$

$$[-4.5; 31.5]$$

Нетипичных значений нет, все входят в границы типичных значений.

d. «Ящик с усами» и гистограмма:



e. Выборочное среднее:  $\bar{x} = 11.2$  (для удобства дальнейших расчетов округлим до 11)

$$\text{Выборочная дисперсия: } s^2 = \frac{(4 - 11)^2 + (18 - 11)^2 + \dots + (9 - 11)^2}{9 - 1} \approx 63$$

$$\text{Выборочное стандартное отклонение: } s = \sqrt{63} \approx 7.9$$

**Задача 2.** Дана выдача R, в которой представлено описание переменной *число лошадиных сил автомобиля*:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
52.0	96.5	123.0	146.7	180.0	335.0

Поясните, что представляет собой каждое значение в выдаче.

**Решение.**

Самое маленькое число лошадиных сил автомобиля в выборке равно 52, самое большое – 335. Среднее число лошадиных сил автомобиля в выборке равно 146.7. Половина автомобилей в выборке имеют число лошадиных сил не выше 123. 25% автомобилей в выборке имеют число лошадиных сил не выше 96.5. 75% автомобилей в выборке имеют число лошадиных сил не выше 180 (или 25% автомобилей – от 180 и выше).

**Задача 3.** Представьте, что вам необходимо построить доверительный интервал для средней стоимости квартиры (в рублях) в Центральном округе Москвы. Известно, что доверительный интервал будет строиться на основе стоимости 26 квартир Центрального округа.

Какое распределение имеет статистика, на значение которой будет умножаться стандартная ошибка выборки при расчете доверительного интервала? Найдите значение этой статистики, приняв уровень доверия равным: а) 90%; б) 95%; в) 99%; г) 99.5%.

**Решение.**

Статистика имеет распределение Стьюдента (t-распределение).

Распределение Стьюдента с числом свободы  $df = n - 1 = 26 - 1 = 25^i$ .

Если выбрать уровень доверия  $\gamma = 0.9$ , то тогда  $t^*$  – это квантиль уровня 0.95, и тогда  $t^* = t_{0.95, df=25} = 1.711$ .

Если выбрать уровень доверия  $\gamma = 0.95$ , то тогда  $t^*$  – это квантиль уровня 0.975, и тогда  $t^* = t_{0.975, df=25} = 2.064$ .

Если выбрать уровень доверия  $\gamma = 0.99$ , то тогда  $t^*$  – это квантиль уровня 0.995, и тогда  $t^* = t_{0.995, df=25} = 2.797$ .

Если выбрать уровень доверия  $\gamma = 0.995$ , то тогда  $t^*$  – это квантиль уровня 0.9975, и тогда  $t^* = t_{0.9975, df=25} = 3.745^{ii}$ .

**Задача 4.** Исследователь оценивает долю жителей страны А, которые поддерживают новый закон о налогообложении. Он опросил 1600 человек и выяснил, что 1000 человек из них поддерживают новый закон. Постройте 92%-ный доверительный интервал для доли жителей, поддерживающих новый закон. Что произойдет с длиной доверительного интервала, если увеличить объем выборки в 4 раза (при прочих равных условиях)?

**Решение.** Выборочная доля:  $\hat{p} = 1000/1600 = 0.625 \approx 0.63$

Стандартная ошибка  $se = \sqrt{\frac{\hat{p}\hat{q}}{n}} \approx 0.012$

Значение z-статистики  $z^* = z_{0.96} = 1.75$

Доверительный интервал:  $[0.63 - 1.75 \cdot 0.012; 0.63 + 1.75 \cdot 0.012]$ ,  $[0.61; 0.65]$ .

Если увеличить объем выборки в 4 раза, то длина доверительного интервала уменьшится в два раза.

**Задача 5.** Группа политологов проводит исследование, посвященное политическим предпочтениям молодежи. Выяснилось, что 62% респондентов придерживаются либеральных взглядов. Вам необходимо проверить гипотезу о равенстве доли приверженцев либеральных взглядов 0.55.

а. Сформулируйте нулевую гипотезу. Сформулируйте альтернативную гипотезу, считая, что она двусторонняя.

<sup>i</sup> Данное значение отсутствует в таблице, для расчетов берем ближайшее  $df = 24$  (можно 26, главное указать). В экзамене будут значения, которые есть в таблице.

<sup>ii</sup> Тоже берем ближайшее – с уровнем 0.9995

- b. Какое распределение имеет статистика критерия, используемого для проверки нулевой гипотезы?
- c. Наблюдаемое значение статистики критерия равно 2.19. Проверьте гипотезу на 5% уровне значимости, построив критическую область для значений критерия. Сделайте статистический и содержательный вывод.
- d. Рассчитайте  $p$ -value. Что означает это значение? Есть ли основания отвергнуть нулевую гипотезу?
- e. Изменится ли вывод относительно нулевой гипотезы, если мы примем уровень значимости равный 10%? Уровень значимости равный 1%?

**Решение.**

- a.  $H_0 : p = 0.55, H_1 : p \neq 0.55$
- b. Статистика имеет стандартное нормальное распределение.
- c.  $z_{набл} = 2.19$ . Критические значения для  $\alpha = 0.05$  с учетом двусторонней альтернативы:  $z_{крит} = 1.96$  и  $-z_{крит} = -1.96$ . Критическая область: значения  $z$  менее  $-1.96$  и более  $1.96$ .  $z_{набл} > 1.96$ , попадает в критическую область, поэтому нулевую гипотезу следует отвергнуть. Статистический вывод: на имеющихся данных на уровне значимости 5% есть основания отвергнуть нулевую гипотезу в пользу альтернативы. Содержательный вывод: истинная доля приверженцев либеральных взглядов среди молодежи не равна 0.55.
- d.  $p$ -value =  $P(z > |z_{набл}|) = 2 \cdot P(z > 2.19) = 2 \cdot (1 - P(z < 2.19)) = 2 \cdot (1 - 0.9857) = 2 \cdot 0.0143 = 0.0286$ . Вероятность того, что мы получим значение  $z$  равное 2.19 или более нетипичное при условии, что нулевая гипотеза верна. Так как  $0.0286 < 0.05$ , на 5% уровне значимости на имеющихся данных есть основания отвергнуть нулевую гипотезу.
- e. Если возьмем уровень значимости  $\alpha = 0.1$ , то вывод относительно нулевой гипотезы не изменится. Если выберем  $\alpha = 0.01$ , то вывод изменится,  $p$ -value  $> \alpha$ , поэтому на 1% уровне значимости гипотеза не отвергается.

**Задача 6.** Проводится исследование, посвященное уровню жизни в регионах Российской Федерации. Разработана методика оценки, сконструирован интегральный индекс уровня жизни. Проверяется гипотеза о равенстве индекса уровня жизни в регионах европейской части России и азиатской части России. В исследовании изучается 18 регионов из европейской части и 14 регионов из азиатской части России.

- a. Сформулируйте нулевую гипотезу и правостороннюю альтернативную гипотезу.
- b. Какое распределение имеет статистика критерия, используемого для проверки нулевой гипотезы? Постройте критическую область для статистики критерия, считая, что нулевая гипотеза проверяется на 10% уровне значимости.
- c. Выдача R по результатам проверки нулевой гипотезы выглядит так:

```
data: e and a
t = 7.9266, df = 138.93, p-value = 6.613e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.47209 10.77389
sample estimates:
mean of x mean of y
74.12198 65.49899
```

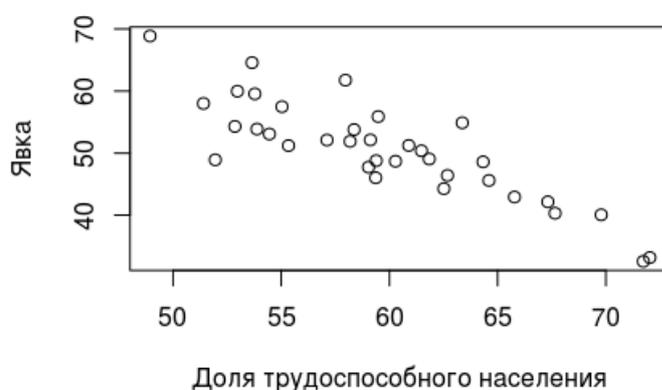
Какой вывод, касающийся нулевой гипотезы, можно сделать, если мы проверяем ее на 10% уровне значимости? Сделайте содержательный вывод.

**Решение.** а.  $H_0 : a_1 = a_2$ ,  $H_1 : a_1 > a_2$

б. Распределение Стьюдента (t-распределение) с  $df = 18 + 14 - 2 = 30$  (в R считается несколько иначе). Критическое значение:  $t_{крит} = t_{0.9, df=30} = 1.31$ . Критическая область: значения  $t$  от 1.31 и более.

с. На имеющихся данных на 10% уровне значимости есть основания отвергнуть нулевую гипотезу о равенстве средних ( $p\text{-value} < \alpha$ ). Содержательный вывод: среднее значение уровня жизни в европейских регионах выше, чем в азиатских.

**Задача 7.** Дана диаграмма рассеяния явки на выборы (в процентах) и доли трудоспособного населения (в процентах):



- Какой вывод о связи двух показателей можно сделать? Можно ли по этому графику сделать вывод, что процент трудоспособного населения оказывает отрицательное влияние на явку?
- Представьте, что вам необходимо проверить статистически, есть ли связь между этими двумя показателями. Вы посчитали коэффициент корреляции Пирсона. Сформулируйте нулевую гипотезу, которую необходимо проверить, чтобы выяснить, значим ли коэффициент корреляции. Какое распределение имеет статистика критерия, используемого для проверки гипотезы?
- Какое значение должно быть у  $p\text{-value}$ , чтобы коэффициент корреляции считался значимым на 5% уровне значимости?

**Решение.** а. Связь обратная, достаточно сильная – точки образуют «облако», имеющее отрицательный наклон. Нет, нельзя, так как диаграмма рассеяния не показывает причинно-следственные связи.

б.  $H_0 : \rho = 0$ ,  $H_1 : \rho \neq 0$

с. Чтобы коэффициент считался значимым, нужно, чтобы гипотеза о равенстве коэффициента нулю была отвергнута. В таком случае  $p\text{-value}$  должен быть менее уровня значимости, то есть менее 0.05.

**Задача 8.** Два эксперта оценивают эссе абитуриентов по английскому языку. Оценки выставляются в 10-балльной шкале. Перед вами оценки, поставленные шести абитуриентам:

Эксперт 1: 10, 8, 6, 4, 3, 9

Эксперт 2: 8, 9, 3, 5, 4, 10

- a. Рассчитайте коэффициент корреляции Спирмена между оценками, поставленными двумя экспертами.
- b. Проверьте гипотезу о равенстве коэффициента корреляции нулю. Сделайте статистический и содержательный вывод. Можно ли считать, что эксперты ставят оценки согласованно?

**Решение.**

- a. Расчет:

$X$	$Y$	$K$	$L$	$d_i = K - L$	$d_i^2$
10	8	6	4	2	4
8	9	4	5	-1	1
6	3	3	1	2	4
4	5	2	3	-1	1
3	4	1	2	-1	1
9	10	5	6	-1	1

$$\sum d_i^2 = 12$$

$$R = 1 - \frac{6 \cdot 12}{6(6^2 - 1)} = 1 - 0.33 = 0.67$$

Прямая согласованность рангов, связь умеренная (ближе к сильной).

- b.  $H_0 : \rho = 0, H_1 : \rho \neq 0$

$$z_{набл} = R \cdot \sqrt{n - 1} = R \cdot \sqrt{6 - 1} = 0.67 \cdot \sqrt{5} \approx 1.83$$

Примем уровень значимости равный 5%.

$$z_{крит} = z_{0.975} = 1.96$$

$$z_{набл} < z_{крит} \Rightarrow$$

Нулевая гипотеза не отвергается, коэффициент равен 0, связи нет. Казалось бы, эксперты ставят оценки согласованно: они ставят высокие оценки одним и тем же абитуриентам, и наоборот. Но: коэффициент незначим, поэтому такой вывод о согласованности сделать нельзя.

**Задача 9.** Дана таблица сопряженности двух признаков: *пол респондента* и его ответ на вопрос: «Любите ли вы горький шоколад?».

	Да	Нет
Женский	32	14
Мужской	15	25

- a. Вам необходимо проверить, есть ли связь между этими признаками. Сформулируйте нулевую и альтернативную гипотезу. Какое распределение имеет статистика критерия, используемого для проверки нулевой гипотезы?
- b. Известно, что p-value равно 0.006. Какой вывод относительно нулевой гипотезы можно сделать?

**Решение.**

$H_0$  : признаки независимы.  $H_1$  : признаки не независимы (связаны). Распределение хи-квадрат. На любом уровне значимости (1%, 5%, 10%) нулевую гипотезу следует отвергнуть, так как p-value («жизнеспособность» нулевой гипотезы) всегда меньше уровня значимости (0.01, 0.05, 0.1). Признаки связаны.

**Задача 10.** Известно, что коэффициент корреляции Пирсона между индексом политических свобод (принимает значения от 1 до 100) и ВВП на душу населения равен 0.76. Проверьте на 5% уровне значимости гипотезу о равенстве истинного значения коэффициента корреляции нулю, построив критическую область для значений статистики используемого критерия. Объем выборки равен 24.

**Решение.**

$$t_{\text{набл}} = \frac{R}{\sqrt{1-R^2}} \cdot \sqrt{n-2} = \frac{0.76}{\sqrt{1-0.76^2}} \cdot \sqrt{22} \approx 5.48$$

$$H_0 : \rho = 0, H_1 : \rho \neq 0$$

$$t_{\text{крит}} = t_{0.975, df=24-2} = 2.074$$

$$t_{\text{набл}} > t_{\text{крит}} \Rightarrow$$

нулевая гипотеза отвергается, связь есть.

**Задача 11.** Найдите:

a.  $P(Z < 1.12), P(Z > 2.32), P(Z < 4.82), P(Z < -0.89)$ .

b.  $P(T < 0), P(T > 0)$ .

где  $Z$  – стандартная нормальная величина,  $T$  – величина, имеющая распределение Стьюдента с  $df = 12$ .

**Решение.**

a. 0.8686, 0.0102, 1, 0.1867

b. 0.5, 0.5

**Задача 12.** Известно, что оценки студентов за тест по курсу «Категории политической науки» имеют нормальное распределение со средним значением 6 и дисперсией 4. Из всех студентов независимо выбирают 100 выборок по 100 человек в каждой. Какое распределение будет иметь набор средних значений, посчитанных по этой выборке? С какими параметрами?

**Решение.** Набор средних значений будет иметь нормальное распределение со средним значением 6 и дисперсией 4/100 (стандартным отклонением 2/10).

**Задача 13.** Сотрудники социологического агенства «НЕ (В)ОПРОС» по результатам опроса выяснили, что 87% респондентов верят в то, что если дорогу перебежит черная кошка – это к неудаче. Найдите вероятность того, что в выборке из 1200 респондентов будет от 600 до 1000 человек, которые верят в эту примету.

**Решение.**  $p = 0.87, q = 0.13, np = 1044, \sqrt{npq} \approx 12$  Используем теорему Муавра-Лапласа.

$$P(600 < S < 1000) = P\left(\frac{600 - 1044}{12} < Z < \frac{1000 - 1044}{12}\right) = P(-37 < Z < -3.6) = \Phi(37) - \Phi(3.6) \approx 0$$