УЧЕБНОЕ ПОСОБИЕ

1919

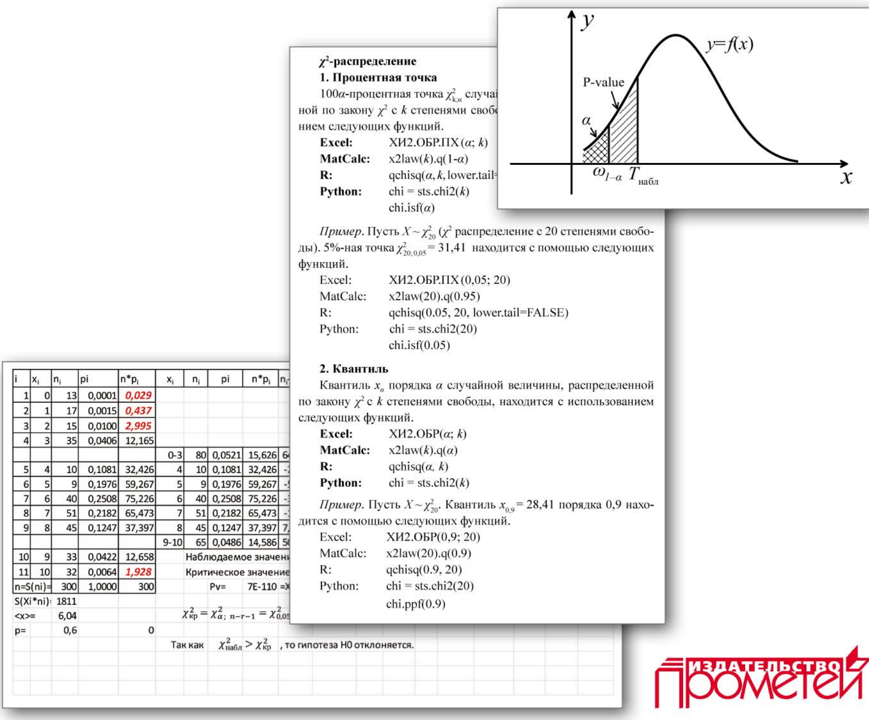
100 ЛЕТ

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ

В. И. Глебов
С. Я. Криволапов

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Проверка гипотез с использованием Excel, MatCalc, R и Python



ИЗДАТЕЛЬСТВО ПРОМЕТЕЙ

Федеральное государственное образовательное бюджетное учреждение высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(ФИНАНСОВЫЙ УНИВЕРСИТЕТ)
Департамент анализа данных, принятия решений и финансовых технологий

В. И. Глебов, С. Я. Криволапов

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Проверка гипотез с использованием Excel, MatCalc, R и Python

Учебное пособие

Предназначено для студентов, обучающихся по направлениям «Экономика», «Менеджмент» и др. Содержит набор задач, посвященных теме «Проверка статистических гипотез» и примеры их решения с использованием Excel, MatCalc, R и Python.

ИЗДАТЕЛЬСТВО ПРОМЕТЕЙ

Москва 2019

УДК 519.2(075.8)
ББК 22.17я73
Г 53

Рецензенты:

Рябов П.Е., д.ф.-м.н., профессор Департамента анализа данных, принятия решений и финансовых технологий Финансового университета при правительстве Российской Федерации.

Спиридонова Т.А., профессор кафедры информатики и математики Всероссийской академии внешней торговли Минэкономразвития РФ.

Глебов В.И., Криволапов С.Я.

Г 53 **Практикум по математической статистике. Проверка гипотез с использованием Excel, MatCalc, R и Python:** Учебное пособие / В.И. Глебов, С.Я. Криволапов. — М.: Прометей, 2019. — 86 с.

ISBN 978-5-907100-66-4

В пособии приведено около 100 задач, посвященных проверке статистических гипотез. Представлены темы: 1) общие свойства (статистика критерия, мощность критерия, P-значение); 2) критерии случайности, независимости, однородности; 3) критерии согласия (простые гипотезы); 4) критерии согласия (сложные гипотезы); 5) параметрические гипотезы.

Разобраны примеры решения задач проверки гипотез. При решении примеров использовалась среда Excel, языки MatCalc [1, 6], R [3] и Python [4].

Пособие предназначено для студентов, обучающихся по направлениям «Экономика», «Менеджмент» и др.

УДК 519.2(075.8)
ББК 22.17я73

ISBN 978-5-907100-66-4

Содержание

Используемые обозначения [2, 5]

$F(x) = F_X(x) = P\{X \le x\}$ — функция распределения случайной величины X

$f(x) = f_X(x)$ — плотность распределения случайной величины X

$E(X)$ — математическое ожидание случайной величины X

$Var(X)$, σ_x^2, $D(X)$ — дисперсия случайной величины X

σ_x — стандартное (среднее квадратическое) отклонение случайной величины X

x_α — квантиль порядка α случайной величины

$\omega_\alpha = x_{1-\alpha}$ — (верхняя) 100α-процентная точка случайной величины

Bin(n; p) — случайная величина, распределенная по биномиальному закону с параметрами n и p

П(λ), Pois(λ) — случайная величина, распределенная по закону Пуассона с параметром λ

Geom(p) — случайная величина, распределенная по геометрическому закону с параметром p $(P\{X = k\} = (1-p)^{k-1}p)$

HGeom(K; L; l) — случайная величина, распределенная по гипергеометрическому закону с параметрами K, L и l $\left(P\{X = k\} = \dfrac{C_K^k C_{L-K}^{l-k}}{C_L^l} \right)$

Unif(a; b), U(a; b) — случайная величина, распределенная по равномерному закону на отрезке [a; b]

Exp(λ) — случайная величина, распределенная по показательному (экспоненциальному) закону с параметром λ

N(m; σ) — нормально распределенная случайная величина с параметрами m и σ (с математическим ожиданием m и дисперсией σ^2)

LN(m; σ) — случайная величина, распределенная по логарифмически нормальному закону с параметрами m и σ

T_k — случайная величина, распределенная по закону Стьюдента (t-распределение) с k степенями свободы

χ_k^2 — случайная величина, распределенная по закону χ^2 с k степенями свободы

$F_{k_1;k_2}$ — случайная величина, распределенная по закону Фишера (F-распределение) с k_1 и k_2 степенями свободы

$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_0^x e^{-\frac{t^2}{2}}dt$ — функция Лапласа

$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ — функция Гаусса

$F_{m,\sigma}(x)=\frac{1}{\sqrt{2\pi}\sigma}\int\limits_{-\infty}^x e^{-(t-m)^2/2\sigma^2}dt=\Phi\left(\frac{x-m}{\sigma}\right)+0,5$ — функция распределения случайной величины $X\sim \mathrm{N}(m;\sigma)$

$f_{m,\sigma}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-m)^2/2\sigma^2}=\frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right)$ — плотность распределения случайной величины $X\sim \mathrm{N}(m;\sigma)$

$\overline{x}$ — выборочное среднее

$\hat{\sigma}_X^2$ — выборочная дисперсия

s_X^2 — исправленная (уточненная) выборочная дисперсия

$\mathrm{R}(X,Y)$, ρ_{xy} — выборочный коэффициент корреляции случайных величин X и Y

z_α — 100α-процентная точка стандартного нормального распределения

$t_{k;\alpha}$ — 100α-процентная точка распределения Стьюдента с k степенями свободы

$\chi_{k;\alpha}^2$ — 100α-процентная точка χ^2-распределения с k степенями свободы

$f_{k_1;k_2;\alpha}$ — 100α-процентная точка F-распределения с k_1 и k_2 степенями свободы

1. Проверка гипотез. Общие свойства

1. Поставщик подготовил партию товаров большого объема для передачи ее покупателю. Данная партия может содержать долю дефектных изделий. Поставщик полагает, что эта доля составляет 3 %, а покупатель — 10 %. Условия поставки: если при проверке 20 случайным образом отобранных товаров обнаружено не более одного дефектного, то партия принимается на условиях поставщика, в противном случае — на условиях покупателя. Требуется определить:

1) каковы статистические гипотезы, статистика критерия, область ее значений, критическая область;

2) какой тип распределения имеет статистика критерия, в чем состоят ошибки первого и второго рода и каковы их вероятности.

При решении задачи используйте тот факт, что гипергеометрическое распределение $\mathrm{HGeom}(N; M; n)$ при малых значениях отношения n/N может быть приближенно заменено биномиальным распределением $\mathrm{Bin}(n; p)$ с параметром $p = M/N$.

2. Фирма производит оковалки. Если производство в порядке, то известно, что 10 % оковалков дефектные. Если производственный процесс нарушен, то 40 % оковалков дефектные. Отдел технического контроля тестирует четыре оковалка каждый час. Если два или больше из них дефектные, то производство останавливается и производится наладка оборудования.

(а) Сформулируйте ситуации ошибок первого и второго рода;

(б) Какова вероятность того, что производство будет остановлено, если оборудование в полном порядке («ложная тревога»)?

(в) Какова вероятность того, что не будет обнаружено нарушение производственного процесса («пропущенная ошибка»);

(г) Предположим, что имеется дополнительная информация:

1) производственный процесс нарушен примерно в 10 % случаев;

2) тестирование одного оковалка стоит 10 у. е.;

3) «пропущенная ошибка» стоит 10000 у. е.;

4) «ложная тревога» стоит 2000 у. е.

Рассчитайте ожидаемые потери, связанные с нарушением производственного процесса. Будет ли лучше использовать другое решающее правило (номер 2): останавливать производство, если по крайней мере один из четырех тестируемых модулей дефектный?

(д) новый менеджер предложил тестировать 100 оковалков и останавливать производство, если оказалось более чем 25 дефектных оковалков. Рассчитайте ожидаемые потери при применении этого решающего правила (номер 3). Сравните новое правило с первыми двумя.

3. Стрелок по летающим тарелкам попадал в цель в 80% случаев. После тренировок в учебном центре он стал поражать мишени в 89 случаях из 100.

а) Можно ли на уровне значимости $\alpha = 0{,}05$ считать, что его квалификация улучшилась?

б) Найдите Р-значение критерия.

Для проверки гипотезы $H_0 : p = p_0$ против гипотезы $H_1 : p > p_0$ используется статистика $Z = \dfrac{\left(\dfrac{k}{n} - p_0\right)}{\sqrt{p_0\left(1 - p_0\right)/n}}$ с правосторонней критической областью $K = \left\{\vec{x} : Z\left(\vec{x}\right) > Z_{\text{кр}}\right\}$, где $Z_{\text{кр}} = z_\alpha$ (z_α — 100α-процентная точка стандартного нормального распределения). При больших значениях параметра n, закон распределения статистики Z имеет приближенно стандартное нормальное распределение N(0; 1).

4. Случайная величина X имеет нормальное распределение с неизвестным параметром m и известной дисперсией $\sigma^2 = 25$. Размер выборки: $n = 100$. Проверяется нулевая гипотеза $H_0 :\ m = m_0 = 50$ против альтернативной гипотезы $H_1 :\ m = 48$. Уровень значимости: $\alpha = 0{,}1$. Статистика критерия имеет вид: $Z = \dfrac{\overline{X} - m_0}{\sigma / \sqrt{n}} = \dfrac{\overline{X} - 50}{5/10}$. Нулевая гипотеза отклоняется, если $Z_{набл} < -z_\alpha$. Найдите мощность критерия w.

При решении задачи используйте тот факт, что, при условии справедливости гипотезы H_0 статистика Z имеет стандартный нормальный закон распределения N(0; 1).

5. Случайная величина X имеет нормальное распределение $\mathrm{N}(m, \sigma)$. Пусть $m = 0$, а размер выборки: $n = 20$. Проверяется нулевая гипотеза $H_0:\ \sigma^2 = \sigma_0^2 = 1$ против альтернативной гипотезы $H_1:\ \sigma^2 > 1$. Уровень значимости: $\alpha = 0{,}05$. Статистика критерия имеет вид:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = (n-1)S^2, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - m\right)^2 .$$

Нулевая гипотеза отклоняется, если $\chi^2_{набл} > \chi^2_{\alpha;n}$ ($\chi^2_{\alpha;n}$ — процентная точка χ^2-распределения с n степенями свободы порядка α). Постройте график функции мощности критерия для значений $w(\sigma^2)$ для значений $\sigma^2 \in [1;4]$.

При решении задачи используйте тот факт, что, при условии справедливости гипотезы H_0 статистика χ^2 имеет χ^2-распределение с n степенями свободы.

6. В лототрон запускаются три шарика с номерами 1, 2 и 3. Вероятность выпадения шарика с номером 2 равна р. Проверяется гипотеза $H_0: p = p_0 = 1/3$, альтернативная гипотеза $H_0: p \neq p_0 = 1/3$. Лототрон запускается $N = 8$ раз.

Критерий: если 2-й шарик выпадает от $k_1 = 3$ до $k_2 = 5$ раз, гипотеза H_0 принимается, в противном случае отклоняется.

а) Найдите величину α — вероятности ошибки первого рода.

б) Постройте график функции мощности $w(p)$ выбранного критерия.

в) Найдите значение мощности критерия при $p = p = 1/3$, $p = p_1 = 0.2$ и $p = p_2 = 0.6$.

г) Повторить то же самое для случая $k_1 = 1$; $k_2 = 4$.

7. Пусть $X_1, \ldots, X_{10}$ выборка из распределения Пуассона с параметром λ. Рассматриваются две простые гипотезы: $H_0 : \lambda = 1$ и $H_1 : \lambda = 2$. Критерий предписывает принимать первую гипотезу H_0, если $X_{(10)} \leq 1$, и альтернативу H_1 в противном случае. Найдите уровень значимости и мощность этого критерия.

8. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения со средним m и единичной дисперсией. Для проверки гипотезы $H_0 : m = 0$ против альтернативы $H_1 : m = 1$ используется следующий критерий: нулевая гипотеза принимается, если $X_{(n)} < 3$ и отклоняется в противном случае. Найдите вероятности ошибок первого и второго рода.

9. Фирма — изготовитель женских украшений, выпустив новый товар, утверждает, что 40% покупателей купят эти украшения. В ходе 10-дневной рекламной распродажи в среднем приобрели украшения 29,5% покупателей, уточненное среднее квадратичное отклонение s составило 16,5%.

а) Может ли оказаться верным утверждение изготовителя товара на 5%-м уровне значимости?

б) Найдите P-значение критерия.

Предполагается, что количество покупателей можно описать нормальным законом распределения.

Для проверки гипотезы $H_0 : m = m_0$ против гипотезы $H_1 : m < m_0$ используется статистика $T = \dfrac{\overline{X} - m_0}{S / \sqrt{n}}$ с левосторонней критической областью $K = \left\{ \vec{x} : T\left(\vec{x}\right) < T_{\text{кр}} \right\}$, где $T_{\text{кр}} = -t_{\alpha;n-1}$ ($t_{\alpha;n-1}$ — процентная точка распределения Стьюдента с $n - 1$ степенью свободы порядка α).

При решении задачи используйте тот факт, что при условии справедливости гипотезы H_0 статистика T имеет закон распределения Стьюдента с $n - 1$ степенями свободы.

10. Имеется игральный кубик с вероятностью выпадения цифры 6, равной p. Проверяется гипотеза $H_0 : p = p_0 = \dfrac{1}{6}$, альтернативная гипотеза $H_1 : p \neq \dfrac{1}{6}$. Кубик подбрасывается 10 раз. Если цифра 6 выпадает 1 или 2 раза, гипотеза H_0 принимается, в противном случае отклоняется.

(а) Найдите α — вероятность ошибки первого рода.

(б) Постройте график функции мощности $w(p)$.

(в) Найдите значение мощности критерия при $p = \dfrac{1}{7}$ и $p = \dfrac{1}{3}$.

11. Случайная величина X имеет нормальное распределение $\mathrm{N}(m;\sigma)$. Дисперсия известна: $\sigma^2 = 4$. По выборке объема $n = 100$ получено выборочное среднее $\overline{x} = 0$. Проверяется нулевая гипотеза $H_0 : m = m_0 = 0$ против альтернативной гипотезы $H_1 : m > 0$. Уровень значимости: $\alpha = 0{,}025$. Статистика критерия имеет вид: $Z = \dfrac{\overline{X} - m_0}{\sigma / \sqrt{n}} = \dfrac{\overline{X}}{2/10}$. Нулевая гипотеза отклоняется, если $Z_{набл} > z_\alpha$. Постройте график функции мощности критерия $w(m)$, для значений $m \in [0;\ 1{,}5]$.

При решении задачи используйте тот факт, что, при условии справедливости гипотезы H_0 статистика Z имеет стандартный нормальный закон распределения $\mathrm{N}(0;1)$.

12. Случайная величина X имеет показательное распределение с параметром λ.

а) Постройте критерий проверки гипотезы $H_0 : \lambda = \lambda_0$ против альтернативы $H_1 : \lambda \neq \lambda_0$, где λ_0 — некоторое заданное число. Для построения критерия используйте следующее утверждение: статистика $T = 2\lambda n\overline{X}$, где выборочное среднее $\overline{X}$ вычислено по выборке объема n из показательного распределения с параметром λ, имеет χ^2-распределение с $2n$ степенями свободы.

б) Используя построенный критерий, проверьте на уровнях значимости $\alpha = 0{,}01$, $0{,}05$ и $0{,}10$ гипотезу $H_0 : \lambda = \lambda_0 = 0{,}2$ против альтернативы $H_1 : \lambda = \lambda_1 \neq \lambda_0$, если по выборке объема $n = 100$ найдено выборочное среднее $\overline{x} = 4{,}5$.

в) Для трёх значений уровня значимости постройте функцию мощности критерия $w(\lambda_1;\ \alpha)$ против гипотез $H_1 : \lambda = \lambda_1 \neq \lambda_0$ для значений $\lambda \in [0{,}1;\ 0{,}4]$ и найдите мощности для случая, когда $H_1 : \lambda = \lambda_1 = 0{,}15$.

г) Найдите P-значение.

13. Автоинспекция обеспокоена превышением скорости движения на участке шоссе с разрешенным движением не более $m_0 = 70$ км/ч. Замер скорости семи случайно выбранных автомобилей радаром дал следующие результаты (км/ч):

79 73 68 77 86 71 69

(а) Дают ли указанные данные основания для вывода о превышении средней скорости на 1%-м уровне значимости?

(б) Найдите Р-значение критерия.

(в) Есть ли основания для вывода о превышении средней скорости на уровне значимости $\alpha = 0{,}04$?

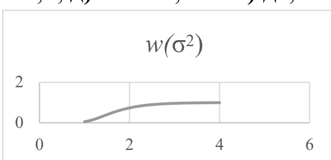
Для проверки гипотезы $H_0 : m = m_0$ против гипотезы $H_1 : m > m_0$ используется статистика $T = \dfrac{\overline{X} - m_0}{S / \sqrt{n}}$ с правосторонней критической областью $K = \left\{\vec{x} : T\left(\vec{x}\right) > T_{\text{кр}}\right\}$, где $T_{\text{кр}} = -t_{\alpha;n-1}$ ($t_{\alpha;n-1}$ — процентная точка распределения Стьюдента с $n - 1$ степенью свободы порядка α).

При решении задачи используйте тот факт, что при условии справедливости гипотезы H_0 статистика T имеет закон распределения Стьюдента с $n - 1$ степенями свободы.

Ответы

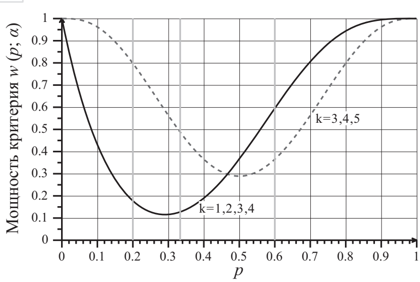
1. 1) $\alpha \approx 0{,}120$; $\beta \approx 0{,}392$. **2. б)** $\alpha \approx 0{,}0523$; **в)** $\beta \approx 0{,}475$; **г)** $\approx 609{,}3$ и $788{,}6$; **д)** $\approx 1001{,}1$. **3. а)** да, можно; **б)** $\approx 0{,}0122$. **4.** $\approx 0{,}767$.

5.



6. а) $\alpha = 0{,}488$

б)

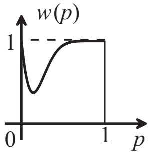


в) $w(1/3) = 0{,}488$; $w(0{,}2) = 0{,}798$; $w(0{,}6) = 0{,}365$;

г) $\alpha = 0{,}127$; $w(1/3) = 0{,}127$; $w(0{,}2) = 0{,}178$; $w(0{,}6) = 0{,}595$.

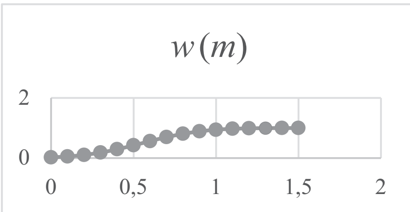
7. $\approx 0{,}99988$. **8.** $\alpha \approx 0{,}0198$; $\beta \approx 0{,}628$. **9. а)** нет, не может; **б)** $\approx 0{,}0375$.

10. а) $\alpha \approx 0{,}386$; **б)**



; **в)** $w(\frac{1}{7}) \approx 0{,}376$; **г)** $w(\frac{1}{3}) \approx 0{,}718$.

11.

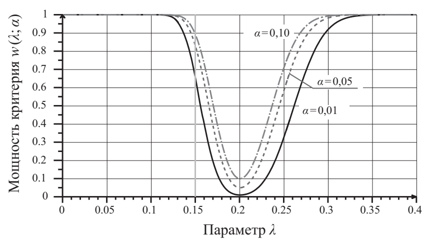


12. а) H_0 принимается, если $T \in \left[\chi^2_{2n;\alpha/2}; \chi^2_{2n;(1-\alpha)/2}\right]$; **б)** H_0 принимается для всех трёх значений α;

в)

α	0,01	0,05	0,10
w	0,6547	0,8315	0,8936

г) P-value = 0,3164.



13. а) да, дают; **б)** $\approx 0{,}0495$; **в)** нет, таких оснований нет.

2. Критерии случайности, независимости, однородности

14. Число e до 20 знака после запятой имеет вид:
2,71828182845904523536.

Используя критерий инверсий, проверьте, обладает ли свойством случайности приведенная последовательность цифр после запятой на уровне значимости $\alpha = 0{,}05$.

15. Сгенерируйте псевдослучайные числа: $m \sim \mathrm{Unif}(-10;\ 10)$, $\sigma \sim \mathrm{Unif}(1;\ 3)$ и $k \sim \mathrm{Unif}(2;\ 2{,}5)$. Используя полученные значения m и σ, сгенерируйте выборку $Y \sim \mathrm{N}(m;\ \sigma)$ объема $n = 200$. Сформируйте выборку X по правилу:

$$X_i = \begin{cases} Y_i, & \text{если} \left|X_i - m\right| < 2{,}25\sigma, \\ k \cdot Y_i, & \text{если} \left|X_i - m\right| \geq 2{,}25\sigma. \end{cases}$$

Проверьте гипотезу о наличии в выборке X аномальных наблюдений.

16. Число $\sqrt{2}$ до 30 знака после запятой имеет вид:
1,414213562373095048801688724209.
Число $\sqrt{3}$ до 30 знака после запятой имеет вид:
1,732050807568877293527446341505.

Используя критерий однородности χ^2 проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что последовательности цифр после запятой для обоих чисел принадлежит одной генеральной совокупности.

17. Имеются данные о значениях роста X (см) и веса Y (кг) 50 выбранных случайным образом мужчин. В результате обработки указанных данных получили следующие выборочные характеристики: $\overline{x} = 171{,}3$; $\overline{y} = 73{,}66$; $\sigma_x^2 = 45{,}33$; $\sigma_y^2 = 39{,}82$; $\overline{xy} = 12645{,}78$. Проверьте на 5%-м уровне значимости гипотезу H_0 об отсутствии значимой линейной корреляционной связи между ростом и весом мужчины.

18. При переписи населения Англии и Уэльса в 1901 г. было зарегистрировано (с точностью до тысячи) 15 729 000 мужчин и 16 799 000 женщин; 3 497 мужчин и 3 072 женщины были зарегистрированы как глухонемые от рождения. Проверьте гипотезу H_0 о том, что глухонемота не связана с полом.

19. Число π до 20 знака после запятой имеет вид:

3,14159265358979323846.

Используя критерий инверсий, проверьте, обладает ли свойством случайности приведенная последовательность цифр после запятой на уровне значимости $\alpha = 0{,}05$.

20. Сгенерируйте выборку Y объема 98 со стандартным нормальным законом распределения $\mathrm{N}(0;1)$. Сформируйте новую выборку X объема $n = 100$, добавив к полученной выборке числа $x_{99} = -3$, $x_{100} = 3{,}5$. Проверьте гипотезу о наличии в выборке X аномальных наблюдений.

21. Число π до 30 знака после запятой имеет вид:

3,141592653589793238462643383279.

Число e до 30 знака после запятой имеет вид:

2,718281828459045235360287471352.

Используя критерий однородности χ^2, проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что последовательности цифр после запятой для обоих чисел принадлежат одной генеральной совокупности.

22. Утверждается, что результат действия лекарства зависит от способа его применения. Проверьте это утверждение на уровне значимости $\alpha = 0{,}05$ по следующим данным:

Результат	A	B	C
Благоприятный	20	23	19
Неблагоприятный	11	17	16

23. Пусть $S(n), n \geq 1$ обозначает цену акции к концу n-й недели, $X = \ln(S(n)/S(n-1))$ — логарифмическая доходность. Изучается связь между логарифмическими доходностями X и Y двух различных биржевых индексов. В этих целях был рассчитан выборочный коэффициент корреляции $\rho_{xy} = 0{,}68$ величин X и Y на основе $n = 62$ выборочных данных. Проверьте на 5 %-м уровне значимости гипотезу H_0 об отсутствии значимой линейной корреляционной связи между доходностями двух биржевых индексов.

24. Средняя температура июня в г. Москве (X) и Ярославле (Y) измерялась в течение 40 лет. Соответствующие данные приведены в следующей таблице:

X	Y	X	Y	X	Y	X	Y	X	Y
12,0	10,0	13,9	11,5	14,9	13,0	16,0	14,8	18,0	16,0
12,0	10,0	13,9	12,0	15,0	13,8	16,8	14,8	18,0	16,0
12,0	10,0	13,9	12,0	15,0	13,9	16,9	15,0	18,1	16,0
12,0	10,1	13,9	12,4	15,0	13,9	16,9	15,0	18,4	16,1
12,8	10,8	13,9	12,9	15,5	14,0	16,9	15,0	19,2	17,0
13,0	10,9	13,9	13,0	15,9	14,2	17,0	15,2	19,3	17,0
13,1	11,0	14,0	13,0	15,9	14,7	17,2	16,0	20,0	17,7
13,8	11,3	14,0	13,0	16,0	14,8	17,5	16,0	20,1	17,8

Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что генеральные совокупности X и Y имеют один и тот же закон распределения.

25. Из 500 абитуриентов, поступавших в институт, 80 человек имели оценку «5» по математике в школе. На вступительном экзамене по математике оценку «5» получили 65 человек. Количество абитуриентов, имевших «5» по математике в школе и сдавших вступительный экзамен по математике на «5», составило 40 человек. С уровнем значимости $\alpha = 0{,}01$ проверьте гипотезу H_0 о независимости оценок «5» в школе и на экзаменах.

26. По статистическим данным рассчитан выборочный коэффициент корреляции между уровнем инфляции X и безработицей Y в некоторой стране за 20 лет: $\tilde{\rho}_{xy} = -0,38$. Проверьте, существует ли значимая корреляционная связь между указанными показателями при уровне значимости $\alpha = 0,05$.

27. Имеется случайная выборка из 10 семей для изучения связи между числом телевизоров (Y) в домохозяйстве и числом членов семьи (X).

X	6	2	4	3	4	4	6	3	2	2
Y	4	1	3	2	2	3	4	1	2	2

Проверьте на уровне значимости $\alpha = 0,05$, существует ли значимая линейная корреляционная связь между случайными величинами X и Y.

28. Изучается зависимость себестоимости единицы продукции (Y, ден. ед.) от величины выпуска продукции (X, тыс. шт.). Были обследованы 5 предприятий и получены следующие данные:

X	2	3	4	5	6
Y	1,9	1,7	1,8	1,6	1,4

Проверьте на уровне значимости $\alpha = 0,05$, существует ли значимая линейная корреляционная связь между случайными величинами X и Y.

29. Сгенерируйте выборку $Y \sim \mathrm{Unif}(0;100)$ объема $n = 200$. Создайте выборку X по следующему правилу: если $Y_i \in (0;95)$, то генерируется $X_i \sim \mathrm{N}(0;1)$, если $Y_i \in (95;100)$, то генерируется $X_i \sim \mathrm{N}(0;4)$ Проверьте на уровне значимости $\alpha = 0,05$ гипотезу о наличии в выборке X аномальных наблюдений.

30. Сгенерируйте четыре независимые выборки: $Y_1 \sim \mathrm{Unif}(-10;10)$ $Y_2 \sim \mathrm{Unif}(-10;-10)$, $Z_1 \sim \mathrm{Exp}(0;1)$, $Z_2 \sim \mathrm{Exp}(0;1)$ (каждую объема 200). Сформируйте выборки X_1, X_2 по правилу: $X_1 = Y_1 + Z_1$, $X_2 = Z_2 - Y_2$. Проверьте на уровне значимости $\alpha = 0,05$

гипотезу $H_0: F_1(x) = F_2(x)$ о совпадении законов распределения случайных величин X_1 и X_2.

31. Сгенерируйте выборки: $Y \sim \mathrm{N}(0; 1)$ и $Z \sim \mathrm{Exp}(0; 1)$ (каждую объема 200). Сформируйте выборки X_1, X_2 по следующему правилу: $X_1 = Y - [Y]$, $X_2 = Z - [Z]$ ($[x]$означает целую часть ($[x] \le x$) числа x). Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0: F_1(x) = F_2(x)$ о совпадении законов распределения случайных величин X_1 и X_2.

32. Сгенерируйте четыре независимые выборки: $Y_1 \sim \mathrm{Unif}(0; 1)$, $Y_2 \sim \mathrm{Unif}(0; 1)$, $Z_1 \sim \mathrm{N}(0; 1)$, $Z_2 \sim \mathrm{N}(0; 1)$ (каждую объема 200). Сформируйте выборки X_1, X_2 по правилу: $X_1 = Y_1^{Z_1}$, $X_2 = Y_2^{Z_2}$. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0: F_1(x) = F_2(x)$ о совпадении законов распределения случайных величин X_1 и X_2.

33. Сгенерируйте выборку $Y \sim \mathrm{Unif}(0; 3)$ объема $n = 200$. Создайте две выборки V и W по следующему правилу: если $Y_1 \in (0; 1)$, то $V_i \sim \mathrm{N}(0; 1)$, $W_i \sim \mathrm{Exp}(1)$; если $Y_i \in (1; 3)$, то $V_i \sim \mathrm{Exp}(1)$, $W_i \sim \mathrm{N}(0; 1)$. Проверьте гипотезу $H_0: F_1(x) = F_2(x)$ о совпадении законов случайных величин V и W.

34. Сгенерируйте две независимые выборки: $Y_1 \sim \mathrm{N}(0; 10)$, $Y_2 \sim \mathrm{N}(0; 10)$ (каждую объема 200). Сформируйте выборки X_1, X_2 по правилу: X_1 равно наибольшему целому числу, меньшему или равному числа Y_1; X_2 равно наименьшему целому числу, большему или равному числа Y_2. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0: F_1(x) = F_2(x)$ о совпадении законов распределения случайных величин X_1 и X_2.

Ответы

14. Обладает. **16.** H_0 принимается. **17.** H_0 отклоняется. **18.** H_0 отклоняется. **19.** Обладает. **21.** H_0 принимается. **22.** Нет, не зависит. **23.** H_0 отклоняется. **24.** H_0 отклоняется. **25.** H_0 отклоняется. **26.** Нет, не существует. **27.** Да, существует. **28.** Да, существует.

3. Критерии согласия. Простые гипотезы

35. В результате 500 подбрасываний монеты орел выпал 270 раз. Проверьте на 5%-м уровне значимости гипотезу H_0 о симметричности монеты.

36. Две монеты подброшены 400 раз. В результате опытов две решки выпали в 90 бросках, орел и решка — в 210 бросках и два орла — в 100 бросках. Проверьте на 5%-м уровне значимости гипотезу H_0 о биномиальном законе распределения числа выпадений орла при двух подбрасываниях монеты.

37. В некоторой стране немцы составляют 50%, французы — 30%, итальянцы — 20%. В гостинице остановились: немцев — 20, французов — 12 и итальянцев — 18. Проверьте на 5%-м уровне значимости гипотезу H_0 о том, что отклонение процентного состава постояльцев по национальностям от среднего по стране объясняется исключительно случайными факторами.

38. Имеется выборка показаний 500 часов, выставленных в витринах часовщиков (час 0 означает промежуток от 0 ч до 1 ч, 1 — от 1 ч до 2 ч и т. д.).

Показания часов	0	1	2	3	4	5	6	7	8	9	10	11
Количество	41	34	54	39	49	45	41	33	37	41	47	39

Выясните, согласуется ли это распределение с теоретическим, имеющим равномерный закон распределения на отрезке $[0; 12]$, на уровне значимости $\alpha = 0{,}05$.

39. Сгенерируйте 50 независимых выборок $Y^{(1)}, Y^{(2)}, \ldots, Y^{(50)}$, каждая из которых представляет собой набор 12 независимых чисел $y_1^{(i)}, y_2^{(i)}, \ldots, y_{12}^{(i)}$ $(i = 1, \ldots, 50)$, равномерно распределенных на отрезке $[0; 1]$. По каждой отдельной выборке сформируйте число x_i по формуле: $x_i = y_1^{(i)} + \ldots + y_{12}^{(i)} - 6$. Проверьте, используя критерий Колмогорова, гипотезу $H_0 : X \sim \mathrm{N}(0; 1)$ о соответствии выборки X стандартному нормальному закону распределения. (Принять уровень значимости $\alpha = 0{,}05$).

40. Сгенерируйте выборку Y объема $n = 200$ с законом распределения: Unif(–1; 1). Сформируйте новую выборку X по правилу: $X = Y^2$. Проверьте, используя подходящие критерии, гипотезу $H_0 : X \sim \text{Unif}(0; 1)$ о соответствии выборки X равномерному на [0; 1] закону распределения. (Принять уровень значимости $\alpha = 0{,}05$).

41. В результате 200 подбрасываний монеты орел выпал 110 раз. Проверьте на 1 %-м уровне значимости гипотезу H_0 о симметричности монеты.

42. В некоторой стране среди всех семей с двумя детьми: семьи с двумя мальчиками составляют 30 %, семьи с двумя девочками — 25 %, семьи с мальчиком и девочкой — 45 %. В одном из регионов страны из 2000 семей с двумя детьми: 500 семей с двумя мальчиками, 600 семей — с двумя девочками и 900 — с мальчиком и девочкой. Проверьте на 5 %-м уровне значимости гипотезу H_0 о том, что отклонение доли семей с разным составом семьи от среднего по стране объясняется исключительно случайными факторами.

43. Сгенерируйте две независимые выборки: $Y \sim \text{Exp}(1)$ и $Z \sim \text{Exp}(1)$, распределенные по показательному закону с параметром $\lambda = 1$ (объема 200). Сформируйте новую выборку X по правилу: $X = Y + Z$. Проверьте, используя подходящие критерии, гипотезу $H_0 : X \sim \chi^2_3$ о соответствии выборки X распределению χ^2 с 3 степенями свободы. (Принять уровень значимости $\alpha = 0{,}05$).

44. Дано распределение числа экзаменов, успешно сданных 50 студентами в сессию (всего было 4 экзамена):

Число успешно сданных экзаменов	0	1	2	3	4
Число студентов	1	1	3	15	30

Проверьте на уровне значимости $\alpha = 0{,}1$ гипотезу H_0 о том, что число сданных экзаменов (из четырех) имеет биномиальный закон распределения.

45. Региональная страховая компания подготовила проект новых схем страхования автомобилей. Случайная выборка 160 владельцев транспортных средств показала, что их мнения по поводу предпочтения предложенных им для рассмотрения четырех схем распределились следующим образом:

Схема	A	B	C	D
Число автомобилистов	34	46	29	51

Проверьте на 5%-м уровне значимости нулевую гипотезу H_0 о том, что владельцы автомобилей в данном регионе не отдают предпочтения ни одной из предложенных схем страхования. Изменится ли ответ, если задан уровень значимости $\alpha = 0{,}04$?

46. В результате проведения 500 опытов получен следующий статистический ряд:

x_i	5	10	20
n_i	280	92	128

Проверьте на 10%-м уровне значимости гипотезу H_0 о том, что случайная величина X распределена по закону, заданному таблицей:

x_i	5	10	20
p_i	0,6	0,15	0,25

47. Среди 10000 «случайных» чисел 0, 1, …, 9, числа, не превосходящие 4, встретились $k = 5089$ раз. Проверьте на уровне значимости $\alpha = 0{,}01$, согласуются ли эти данные с гипотезой H_0 о случайности чисел.

48. Проводились опыты с бросанием одновременно 12 правильных кубиков. Наблюдаемую случайную величину X считали равной числу кубиков, на которых выпало 6 очков.. Пусть n_i — число опытов, в которых наблюдалось значение $X = i$, $i = 0, 1, ..., 12$. Данные для $n = 4096$ опытов приведены в следующей таблице.

x_i	0	1	2	3	4	5	6	≥ 7
n_i	447	1145	1181	796	380	115	24	8

Согласуются ли эти данные на уровне значимости $\alpha = 0{,}1$ с гипотезой H_0 о симметричности кубиков?

49. Сгенерируйте две независимые выборки: $Y_1 \sim \text{Unif}(-1; 1)$, $Y_2 \sim \text{Unif}(-1; 1)$ (каждую объема 200). Сформируйте новые выборки X_1 и X_2 по правилу: $X_1 = Y_1^2$, $X_2 = Y_2^2$. Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы:

1) $H_0 : F_1(x) = F_2(x)$ — о совпадении законов распределения случайных величин X_1 и X_2;

2) $H_0 : X_1 \sim \text{Unif}(0; 1)$ — о соответствии закона распределения случайной величины X_1 равномерному закону $\text{U}(0; 1)$;

3) $H_0 : X_2 \sim \text{Unif}(0; 1)$ — о соответствии закона распределения случайной величины X_2 равномерному закону $\text{U}(0; 1)$.

50. Сгенерируйте псевдослучайное натуральное число k с равномерным на множестве $\{10, 11, ..., 20\}$ законом распределения. Сгенерируйте выборки: $Y \sim \text{Unif}(-k; k)$, $Z \sim \text{Unif}(1; k)$ (каждую объема 200). Сформируйте новую выборку X по правилу: $X = \frac{Y}{Z}$. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о соответствии закона распределения случайной величины X равномерному закону $\text{Unif}(-k; k)$.

51. Сгенерируйте выборку Y объема $n = 200$ со стандартным нормальным законом распределения. Сформируйте новую выборку X по правилу: $X = Y - [Y]$ ($[Y]$ — целая часть числа Y). Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о соответствии закона распределения случайной величины X равномерному закону $\text{Unif}(0; 1)$.

52. Сгенерируйте три независимые выборки Y_1, Y_2 и Y_3 (каждую объема 200) со стандартным нормальным законом распределения. Сформируйте новую выборку X по правилу: $X = Y_1^2 + Y_2^2 + Y_3^2$.

Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы о законе распределения случайной величины X:

1) $H_0 : X \sim \chi_2^2$; 2) $H_0 : X \sim \chi_3^2$; 3) $H_0 : X \sim \chi_2^4$.

53. Сгенерируйте четыре независимые выборки Y_1, Y_2, Y_3 и Y_4, (каждую объема 200) со стандартным нормальным законом распределения. Сформируйте новую выборку X по правилу:

$$X = \frac{Y_1}{\sqrt{\left(Y_2^2 + Y_3^2 + Y_4^2\right)/3}}.$$

Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы о законе распределения случайной величины X:

1) $X \sim \mathrm{N}(0;1)$; 2) $X \sim T_1$; 3) $X \sim T_3$.

54. Сгенерируйте шесть независимых выборок Y_1, Y_2, Y_3, Y_4, Y_5 и Y_6 (каждую объема 200) со стандартным нормальным законом распределения. Сформируйте новую выборку X по правилу:

$$X = \frac{\left(Y_1^2 + Y_2^2\right)/2}{\left(Y_3^2 + Y_4^2 + Y_5^2 + Y_6^2\right)/4}.$$

Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы о законе распределения случайной величины X:

1) $X \sim F_{1;1}$; 2) $X \sim F_{4;2}$; 3) $X \sim F_{2;4}$.

55. Сгенерируйте две независимые выборки Y и Z (каждую объема 200) с законами распределения: $Y \sim \mathrm{Unif}(-1;1)$ и $Z \sim \mathrm{Exp}(0;1)$. Сформируйте новую выборку X по правилу: $X = YZ$.

1) Найдите параметры: $m = E(X)$ и $\sigma^2 = Var(X)$.

2) Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о соответствии закона распределения случайной величины X нормальному закону с параметрами m и σ^2.

56. Сгенерируйте два псевдослучайных числа: $p_0 \sim \mathrm{Unif}(0{,}3;0{,}5)$, $\lambda_0 \sim \mathrm{Unif}(0;10)$. Сгенерируйте две независимые выборки Y и Z (каждую объема 200) с законами распределения: $Y_i \sim \mathrm{Geom}(p_0)$ и $Z_i \sim \Pi(\lambda_0)$. Сформируйте новую выборку X по правилу: $X = YZ$.

1) Найдите параметры: λ = E(X) и p=1/E(X).

2) Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезы:
а) $X \sim \mathrm{Geom}(p)$; б) $X \sim \Pi(\lambda)$.

57. Сгенерируйте псевдослучайные числа: $\lambda \sim \mathrm{Unif}(0;1)$, $m_0 \sim \mathrm{Unif}(0;10)$, и $\sigma_0 \sim \mathrm{Unif}(0;5)$. Сгенерируйте две независимые выборки: $Y_i \sim \mathrm{Exp}(\lambda)$, $Z_i \sim \mathrm{N}(m_0;\sigma_0)$ (каждую объема 200). Сформируйте новую выборку X по правилу: $X = YZ$.

1) Найдите параметры: $m = E(X)$ и $\sigma^2 = D(X)$.

2) Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о соответствии закона распределения случайной величины X нормальному закону с параметрами m и σ.

Ответы

35. H_0 принимается. **36.** H_0 принимается. **37.** H_0 отклоняется. **38.** Согласуется**. 41.** H_0 принимается. **42.** H_0 отклоняется. **43.** H_0 отклоняется. **44.**H_0 принимается. **45.** На 5%-м уровне отклоняется; на 4%-м уровне принимается. **46.** H_0 отклоняется. **47.** H_0 отклоняется. **48.** Согласуется.

4. Критерии согласия. Сложные гипотезы

58. Для каждого из 100 сотрудников компании регистрировалось число обращений в службу ИТ-поддержки в течение года:

Число обращений в службу ИТ-поддержки	0	1	2	3	4 и более
Число компьютеров	50	25	18	7	0

Проверьте на 5%-ном уровне значимости гипотезу о том, что число обращений в службу ИТ-поддержки одним пользователем за год распределено по закону Пуассона.

59. Выборочные данные о значениях случайной величины X заданы в виде интервального ряда

0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50
2	12	8	4	14	6	10	2	1	11

Выясните, согласуется ли это распределение на уровне значимости $\alpha = 0{,}05$ с теоретическим, имеющим равномерный закон распределения.

60. По выборке объема $n = 100$ значений случайной величины X составлен интервальный статистический ряд.

Δ_i	[0; 2)	[2; 4)	[4; 6)	[6; 8)	[8; 10]
n_i	10	25	30	30	5

Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о нормальном законе распределения генеральной совокупности.

61. Десять тысяч странников записались на кастинг для участия в передаче про ясновидящих. Участникам предлагалось «увидеть» по 5 разных предметов, спрятанных за непрозрачным экраном. Список предметов был известен участникам. Во второй строке таблицы показано, сколько участников точно назвали предметы в количестве, указанном в первой строке таблицы.

x_i	0	1	2	3	4	5
n_i	3277	4140	2006	514	61	2

Требуется проверить гипотезу о том, что все участники давали ответы наобум, с вероятностью $p = 1/5$ точно назвать предмет.

При этих предположениях проверьте гипотезу H_0 о биномиальном законе распределения Bin(5;1/5) соответствующей случайной величины на уровне значимости $\alpha = 0{,}05$.

62. Сгенерируйте выборку X объема $n = 200$, каждый элемент которой распределен по показательному закону со случайным параметром λ_i: $\lambda_i \sim \mathrm{Unif}(0; 10)$, $X \sim \mathrm{Exp}(\lambda_i)$.

А. Проверьте, используя подходящие критерии, гипотезу о показательном законе распределения выборки X;

Б. Проверьте, используя подходящие критерии, гипотезу о нормальном законе распределения выборки X.

63. В больнице скорой помощи фиксировалось количество X вызовов в час специализированных бригад. Наблюдения велись в течение 100 часов. Их результаты приведены в таблице.

Число вызовов	0	1	2	3	4	5	6	7
Частота вызовов	6	27	26	20	10	5	5	1

С помощью критерия χ^2 проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0, состоящую в том, что случайная величина X распределена по закону Пуассона.

64. По выборке объема $n = 50$ значений случайной величины X составлен интервальный статистический ряд:

Δ_i	[0; 20)	[20; 40)	[40; 60)	[60; 80)	[80; 100]
n_i	10	10	8	5	7

Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о показательном законе распределения генеральной совокупности.

65. Сгенерируйте две независимые выборки $Y \sim \mathrm{Exp}(1)$, $Z \sim \mathrm{Exp}(1)$ (каждую объема 200), распределенные по показательному закону с параметром $\lambda = 1$. Сформируйте новую выборку X по правилу: $X = Y + Z$. Проверьте, используя подходящие критерии, гипотезу $H_0 : X \sim \mathrm{Exp}(\lambda)$ о соответствии выборки X показательному закону распределения с неизвестным параметром λ. (Принять уровень значимости $\alpha = 0{,}05$).

66. Пользуясь критерием χ^2, установите, на уровне значимости $\alpha = 0{,}05$, согласуется ли с нормальным законом распределение, заданное следующим рядом:

Δ_i	[–2,2; –1,2)	[–1,2; –0,4)	[–0,4; 0,4)	[0,4; 1,2)	[1,2; 2,0]
n_i	6	11	21	7	5

67. Цена X различных типов электроприборов в магазине (в тыс. руб.) представлена в виде статистического ряда

[0–0,9)	[0,9–1,8)	[1,8–2,7)	[2,7–3,6)	[3,6–4,5)	[4,5–5,4}	[5,4–6,3]
25	16	5	1	1	1	1

Проверьте гипотезу H_0 о показательном законе распределения случайной величины X на уровне значимости $\alpha = 0{,}05$.

68. Сгенерируйте две выборки: $Y \sim \mathrm{Exp}(0,1)$, $Z \sim \mathrm{U}(0,1)$ (каждую объема 200). Сформируйте новую выборку X по правилу: $X = Y^Z$. Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы о законе распределения случайной величины X:

1) $H_0 : X \sim \mathrm{Exp}(\lambda)$; 2) $H_0 : X \sim \mathrm{Unif}(a; b)$;

3) $H_0 : X \sim \mathrm{LN}(m; \sigma)$ (λ, a, b, m, σ — неизвестные параметры).

69. Сгенерируйте три псевдослучайных числа: $m_0 \sim \mathrm{Unif}(-1; 1)$, $\sigma_0 \sim \mathrm{Unif}(1; 10)$, $\lambda \sim \mathrm{Unif}(0; 2)$. Сгенерируйте две выборки: $Y_i \sim \mathrm{N}(m_0; \sigma_0)$, $Z \sim \mathrm{Exp}(\lambda)$ (каждую объема 200). Сформируйте новую выборку X по правилу: $X = Y/Z$. Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы о законе распределения случайной величины X:

1) $X \sim N(m_0; \sigma_0)$; 2) $Z_i \sim U(a; b)$, где m, σ, a, b — неизвестные параметры.

70. Сгенерируйте три псевдослучайных числа: $m_0 \sim Unif(-1; 1)$, $\sigma_0 \sim Unif(0; 1)$, $\lambda \sim Unif(0; 2)$. Сгенерируйте две выборки: $Y \sim Exp(\lambda)$, $Z \sim LN(m_0; \sigma_0)$ (каждую объема 200). Сформируйте новую выборку X по правилу: $X = \min(Y; Z)$.

Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о соответствии выборки X нормальному закону $N(m; \sigma)$ с неизвестными параметрами m и σ.

71. Сгенерируйте выборку X объема $n = 200$, каждый элемент которой распределен нормально со случайными параметрами m_i и σ_i^2, где $m_i \sim Unif(-10; 10)$, $\sigma_i \sim Unif(1; 3)$, $X_i \sim N(m_i; \sigma_i)$. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о нормальном законе распределения выборки X.

72. Сгенерируйте выборку X объема $n = 200$, каждый элемент которой распределен равномерно со случайными параметрами a_i и b_i: $a_i \sim Unif(-10; 0)$, $b_i \sim Unif(0; 10)$, $X_i \sim Unif(a_i; b_i)$. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о равномерном законе распределения выборки X.

73. Сгенерируйте выборку $Y \sim Exp(1)$ объема $n = 200$. Найдите реализации случайной величины X, равной наименьшему целому числу, большему или равному числа Y. Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы:

1) $X \sim \Pi(\lambda)$; 2) $X \sim Geom(p)$ (λ, p — неизвестные параметры).

74. Сгенерируйте выборку Y объема $n = 200$, в которой каждый элемент Y_i имеет нормальный закон распределения $Y_i \sim N(m_i; \sigma_i)$ со случайными параметрами m_i и σ_i, и выборку Z, также объема $n = 200$, в которой каждый элемент Z_i распределен по показательному закону распределения $Z_i \sim Exp(\lambda_i)$, где параметр λ_i также является случайным:

$m_i \sim Unif(-10; 10)$, $\sigma_i \sim Unif(1; 3)$, $\lambda_i \sim Unif(0; 1)$, $Y_i \sim N(m_i; \sigma_i)$, $Z_i \sim Exp(\lambda_i)$.

Сформируйте выборку X по правилу: $X = YZ$. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $X \sim \mathrm{N}(m;\ \sigma)$ с неизвестными параметрами m и σ.

75. Сгенерируйте псевдослучайное число n с равномерным законом распределения на множестве $\{1, 2, ..., 10\}$ и выборку l объема $n = 200$, в которой каждый элемент l_i с одинаковой вероятностью 0,5 принимает значения 1 и –1. Сгенерируйте выборку $Y \sim \chi_n^2$. Сформируйте выборку X по правилу: $X = l \cdot \sqrt{Y}$. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : X \sim \mathrm{N}(m; \sigma)$ с неизвестными параметрами m и σ.

76. Сгенерируйте псевдослучайное число n с равномерным законом распределения на множестве $\{1, 2, ..., 10\}$. Сгенерируйте выборку $Y \sim \chi_n^2$ объема $n = 200$. Сформируйте выборку X по правилу: $X = \sqrt{Y}$. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : X \sim \mathrm{Exp}(\lambda)$ с неизвестным параметром λ.

77. Сгенерируйте выборку $Z \sim \mathrm{Unif}(0; 1)$ объема $n = 200$. Сформируйте выборки X и Y по правилу: $X = -\ln Z$, $Y = [X] + 1$. Проверьте на уровне значимости $\alpha = 0{,}05$ следующие гипотезы:

1) $H_0 : X \sim \mathrm{Exp}(\lambda_1)$; 2) $H_0 : X \sim \mathrm{LN}(m; \sigma)$;

3) $H_0 : Y \sim \mathrm{Geom}(p)$; 4) $H_0 : Y \sim \Pi(\lambda_2)$ ($\lambda_1, \lambda_2, p, m, \sigma$ — неизвестные параметры).

Ответы

58. H_0 принимается (P-value = 0,058). **59.** Не согласуется. **60.** H_0 принимается. **61.** H_0 принимается (P-value = 0,808). **63.** H_0 принимается. **64.** H_0 отклоняется. **66.** Не согласуется. **67.** H_0 принимается.

5. Проверка параметрических гипотез

78. Из 200 кредитов, выданных в первом отделении банка, не были возвращены 40, а из 300 кредитов, выданных во втором отделении, не были возвращены 80. Можно ли на этом основании заключить, что вероятность выдачи кредита ненадежному клиенту во втором отделении выше, чем в первом? Проверьте соответствующую гипотезу на 1%-ном уровне значимости.

79. В выборке из десяти студентов время, затраченное на подготовку к экзамену по теории вероятностей равно (в часах):

28; 57; 42; 35; 61; 39; 55; 46; 49; 38.

Проверьте на 5%-ном уровне значимости гипотезу H_0, что среднее время, затраченное на подготовку к экзамену, равно 40 ч.

80. Партия изделий принимается в том случае, если вероятность того, что изделие окажется соответствующим стандарту, будет не менее 0,95. Среди случайно отобранных 100 изделий оказалось 94 соответствующих стандарту.

а) Можно ли на уровне значимости $\alpha = 0{,}1$ принять партию?

б) Найдите P-значение критерия.

81. Многие ботаники делали опыты по скрещиванию желтого (гибридного) гороха. По известной гипотезе Менделя вероятность появления зеленого гороха в таких опытах равна $\frac{1}{4}$. При 34153 опытах скрещивания в 8506 опытах получен зеленый горох. Подтверждают ли эти данные гипотезу Менделя? Уровень значимости принять $\alpha = 0{,}1$.

82. Имеется выборка объема 5 из распределения X:

11; 6; 7; 12; 9.

а) Проверьте гипотезу $H_0 : \sigma^2 = 7$ против альтернативы $H_1 : \sigma^2 \neq 7$ на уровне значимости $\alpha = 0{,}2$.

б) Найдите P-значение критерия.

83. Биржевой маклер исследует возможности двух инвестиций А и В. Инвестиция А предполагается на срок 10 лет с ожидаемой ежегодной прибылью в течение этого периода 17,8%. Инвестиция В рассчитана на срок 8 лет также с ожидаемой ежегодной прибылью 17,8%. Выборочные стандартные отклонения ежегодных прибылей от двух инвестиций составляют 7,14% и 3,21% соответственно. Есть ли основания на 1%-м уровне значимости утверждать, что риски инвестиций А и В не равны?

84. Выборки X и Y представляют собой последовательности 30 знаков после запятой чисел π и e соответственно:

π: 3,141592653589793238462643383279;

e: 2,718281828459045235360287471352.

Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : E(X) = E(Y)$ против альтернативы $H_0 : E(X) \neq E(Y)$.

85. Отдел маркетинга автотранспортного предприятия, занимающегося междугородными перевозками, провел обследование стоимости топлива на бензоколонках на трассе между городами А и В. Результаты показали, что средняя цена одного литра топлива на 52 заправках фирмы «Торопил» — 30,5 руб. с выборочным стандартным отклонением 0,085 руб., а на 58 заправках других фирм средняя цена одного литра — 30,0 руб. с выборочным стандартным отклонением 0,075 руб. Можно ли на уровне значимости $\alpha = 0{,}05$ утверждать, что средняя цена одного литра топлива на заправках «Торопил» существенно выше цены этого же топлива на заправках других фирм?

86. Универсам имеет отдел фасованных продуктов, в котором сахар продается в пакетах весом в 1 кг. Проверяющий случайным образом отобрал 41 пакет и определил, что их средний вес составил 970 г, а исправленное выборочное стандартное отклонение 10 г. Можно ли объяснить такое различие в весе случайностью (при $\alpha = 0{,}05$)?

87. Из партии добытых алмазов случайным образом отобраны 6 экземпляров. Вычисленные по ним выборочный средний вес и выборочное стандартное отклонение составили 0,53 карат и 0,0559 карат соответственно. Проверьте нулевую гипотезу о том, что средний вес алмаза равен 0,5 карат при альтернативной гипотезе о том, что он больше 0,5 карат. Уровень значимости принять равным $\alpha = 0{,}05$.

88. В 2010 г. в Университет поступило 1500 из 4000 юношей, подавших заявление, и 850 из 2500 девушек. Есть ли основания считать, что вероятности поступления в Университет для юношей и девушек отличаются между собой на 1%-м уровне значимости?

89. По выборке объема $n = 10$ найдена исправленная выборочная дисперсия $s^2 = 111$.

а) Проверьте на 5%-м уровне значимости гипотезу $H_0 : \sigma = 9$ против альтернативы $H_1 : \sigma \neq 9$.

б) Найдите P-значение критерия.

90. После выхода негативной информации курс акций компании ABC упал на 10%. До этого по результатам наблюдений за 30 периодов от закрытия до закрытия выборочное стандартное отклонение логарифмической доходности составило $s_x = 1{,}6\%$. В последующие после падения 25 периодов выборочное стандартное отклонение увеличилось до $s_y = 2{,}4\%$. Проверьте на 5%-м уровне значимости нулевую гипотезу о том, что генеральная дисперсия σ^2 не увеличилась. Предполагается, что дневная логдоходность (т.е. логарифмическая доходность от закрытия до следующего ближайшего закрытия торгов) распределена до выхода негативной информации по нормальному закону с параметрами m_x и σ_x^2, а после выхода — с параметрами m_y и σ_y^2.

91. Группа из 20 студентов сдавала экзамены по линейной алгебре и математическому анализу. Пусть X — оценка студента по линейной алгебре, Y — по математическому анализу. Обработка результатов сдачи экзаменов дала следующие результаты: $\overline{x} = 3{,}6$; $\overline{y} = 3{,}4$; $s_x^2 = 0{,}2$; $s_y^2 = 0{,}3$.

а) Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о равенстве средних баллов по одному и другому экзамену.

б) Найдите P-значение критерия.

92. Фирма — производитель топливных добавок утверждает, что их использование позволяет проехать в среднем на 3 км больше на 1 л бензина. Была произведена случайная выборка из 100 автомобилей для проверки этого утверждения. Оказалось, что в среднем пробег на 1 л увеличился на 2,4 км, при этом исправленное выборочное стандартное отклонение составило 1,8 км. Можно ли на 5%-м уровне значимости утверждать, что среднее увеличение пробега составило не менее 3 км?

93. Проверка, проведенная в отделе фасованных продуктов, показала, что средний вес 121 штуки случайно отобранных 60-граммовых пакетиков с маком составил 59 г при исправленном выборочном стандартном отклонении 5 г. Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о том, что полученная разница в весе является случайной против альтернативы: вес пакетиков с маком меньше 60 г.

94. В области лечения онкологических заболеваний была предложена новая методика лечения. Известно, что среди пациентов, которые прошли стандартный курс лечения, 30% прожили более 5 лет после проведенного курса лечения. Курс лечения по новой методике применялся к 100 пациентам, из которых 38 прожили более 5 лет после проведенного курса лечения. Можно ли на уровне значимости $\alpha = 0{,}05$ утверждать, что новая методика лечения более эффективна, чем стандартная?

95. Компания продает автомобили. 13% проданных автомобилей возвращается на исправление недостатков в течение года. Была предпринята кампания по улучшению контроля качества. Через полгода после кампании для проверки ее эффективности была сделана случайная выборка из 1572 проданных машин, из которых вернулись на доработку 133. Была ли кампания успешной? Уровень значимости принять 0,05.

96. Фирма — производитель зубной пасты занимала 23% рынка. Для того чтобы увеличить свою долю рынка, фирма наняла рекламную кампанию. По истечении времени эта кампания сообщила, что после ее усилий из выборки 1000 покупателей 28% выбрали пасту этой фирмы. Свидетельствует ли это о том, что реклама была эффективной? Уровень значимости принять равным 0,05.

97. При 100 бросаниях правильного кубика «шестерка» выпала 20 раз. Согласуется ли этот результат с утверждением, что кубик «правильный» на уровне значимости $\alpha = 0{,}1$?

98. Для оценки доли городских домохозяйств, использующих электронную почту для связи, был проведен опрос 150 случайно выбранных домохозяйств, из которых 45 сообщили, что они используют электронную почту.

Аналогичный опрос был проведен среди 180 сельских домохозяйств, из которых члены 71 домохозяйства сообщили, что они используют электронную почту. Можно ли на основании этих данных утверждать на 5%-м уровне значимости, что жители сельской местности используют электронную почту в большей степени?

99. Важной мерой, ассоциированной с риском акции, является стандартное отклонение или дисперсия движения цены акции. Финансовый аналитик проверяет одностороннюю гипотезу о том, что акция А имеет больший риск (большую вариацию цены), чем акция В. Случайная выборка за 13 дней цены акции А дала величину исправленной выборочной дисперсии, равную $s_x^2 = 6{,}25$; случайная выборка за 18 дней цены акции В дала исправленную выборочную дисперсию $s_y^2 = 3{,}47$. Имеются ли основания на 1%-м уровне значимости считать, что акция А имеет больший риск?

100. Имеются данные о курсах ценных бумаг двух типов X и Y:

x_i	5	6	8	10	11
n_i	10	10	15	5	5

y_i	2	5	7	10	15
n_i	3	12	20	10	5

а) Проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу о равенстве дисперсий двух совокупностей.

б) Найдите Р-значение критерия.

101. В таблице приведены данные по ценам на автомобиль ВАЗ-21102 по выборкам фирм в Москве и Твери ($).

Москва	5770	5810	5700	5725	5650	5760
Тверь	5700	5600	5680	5750	5620	5680

Можно ли, на основании данных таблицы утверждать на уровне значимости $\alpha = 0{,}05$, что цена на автомобиль в Москве выше, чем в Твери?

Ответы

78. а) Да, можно (P-value = 0,0436). **79.** H_0 принимается. **80. а)** Да, можно; **б)** 0,323. **81.** Да, подтверждают. **82. а)** H_0 принимается; **б)** 0,892. **83.** Таких оснований нет. **84.** H_0 принимается. **85.** Да, можно. **86.** Нет, нельзя. **87.** Нет оснований отклонить H_0. **88.** Да, есть. **89. а)** Нет оснований отклонить H_0; **б)** 0,390. **90.** H_0 отклоняется. **91. а)** Нет оснований отклонить H_0; **б)** 0,22. **92.** Нет, нельзя. **93.** H_0 отклоняется. **94.** Да, можно. **95.** Да, была. **96.** Да, свидетельствует. **97.** Да, согласуется. **98.** Да, можно. **99.** Нет, не имеются. **100. а)** H_0 отклоняется; **б)** 0,00224. **101.** Да, можно.

6. Статистические распределения в Excel, MatCalc, R и Python

Ниже перечисляются функции Excel, MatCalc, R и Python, позволяющие вычислять процентные точки, а также значения функций распределения и плотности распределения для некоторых законов распределения.

Если есть намерение использовать статистические функции языка Python, то предварительно должна быть импортирована библиотека scipy.stats. Всюду в дальнейшем предполагается, что выполнена команда импортирования этой библиотеки под именем sts:

import scipy.stats as sts

Нормальный закон распределения

1. Процентная точка.

100α-процентная точка z_α стандартной нормально распределенной случайной величины находится с использованием следующих функций.

Excel: НОРМ.СТ.ОБР$(1-\alpha)$

MatCalk: nlaw(0,1).q$(1-\alpha)$

R: norm($p=\alpha$, mean=0, s = 1, lower.tail=FALSE)

или кратко: qnorm(α, lower.tail=FALSE)

Python: sts.norm.isf(α)

Пример. Значение *5%*-й точки стандартного нормального распределения $z_{0,05} = 1,6449$ находится с использованием следующих функций.

Excel: НОРМ.СТ.ОБР(0,95)

MatCalk: nlaw(0,1).q(0.95)

R: qnorm(0.05, lower.tail=FALSE) (qnorm(0.95, lower.tail=TRUE), или короче, qnorm(0.95))

Python: sts.norm.isf(0.05)

2. Квантиль.

Квантиль x_α порядка α стандартного нормального закона распределения находится с помощью следующих функций.

Excel: НОРМ.СТ.ОБР(α)

MatCalk: nlaw(0,1).q(α)

R: qnorm($p = \alpha$, mean=0, s = 1, lower.tail=FALSE)

или кратко: qnorm(α)

Python: sts.norm.ppf(α)

Пример. Значение квантили порядка 0,9 стандартного нормального закона распределения $x_{0,9}$ = 1,2816 находится с использованием следующих функций.

Excel: НОРМ.СТ.ОБР(0,9)

MatCalk: nlaw(0,1).q(0.9)

R: qnorm(0.9)

Python: sts.norm.ppf(0.9)

3. Функция распределения стандартного нормального распределения $F_{0,1}(x)$

Значение функции распределения стандартного нормального закона $F_{0;1}\left(x\right) = \Phi\left(x\right) + 0{,}5$ в точке x находится с использованием следующих функций.

Excel: НОРМ.СТ.РАСП(x; ИСТИНА)

MatCalk: nlaw(0,1).pl(x)

R: pnorm(x, mean =0, sd = 1)

или кратко: pnorm(x)

Python: sts.norm.cdf(x, loc=0, scale=1)

или кратко: sts.norm.cdf(x)

Пример. Пусть $X \sim \mathrm{N}(0; 1)$. Значение функции распределения $F_{0;1}(1) = 0{,}8413$ в точке $x = 1$ находится с помощью следующих функций.

Excel: НОРМ.СТ.РАСП(1; 1)

MatCalk: nlaw(0,1).pl(1)

R: pnorm(1)

Python: sts.norm.cdf(1)

4. Плотность стандартного нормального распределения $\varphi(x)$

Значение плотности распределения

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

стандартной нормально распределенной случайной величины в точке x вычисляется с использованием следующих функций.

Excel:	НОРМ.СТ.РАСП (x; ЛОЖЬ)
MatCalk:	nlaw(0,1).den(x)
R:	dnorm(x)
Python:	sts.norm.pdf(x)

Пример. Пусть $X \sim \mathrm{N}(0; 1)$. Значение плотности распределения $\varphi(0) = 0{,}3989$ в точке $x = 0$ находится с помощью следующих функций.

Excel:	НОРМ.СТ.РАСП (0; 0)
MatCalc:	nlaw(0,1).den(0)
R:	dnorm(0)
Python:	sts.norm.pdf(0)

5. Функция распределения $F_{\mu;\sigma}(x)$.

Значение функции распределения нормально распределенной случайной величины $X \sim \mathrm{N}(m; \sigma)$ с параметрами m и σ, т.е. значение

$$F_{\mu;\sigma}(x) = \Phi\left(\frac{x-m}{\sigma}\right) + 0{,}5$$

находится с использованием следующих функций.

Excel:	НОРМ.РАСП (x; m; σ; ИСТИНА)
MatCalc:	nlaw(m,σ).pl(x)
R:	pnorm (x, m, σ)
Python:	gaus = sts.norm(loc = m, scale = σ)
	gaus.cdf(x)

Пример. Пусть $X \sim \mathrm{N}(5; 2)$. Значение функции распределения $F_{5;2}(4) = 0{,}3085$ в точке $x = 4$ определяется ется с помощью следующих функций.

Excel: НОРМ.РАСП (4; 5; 2; 1)
MatCalc: nlaw(5,2).pl(4)
R: pnorm(4, 5, 2)
Python: gaus = sts.norm(loc=5, scale=2)
gaus.cdf(4)

6. Плотность распределения $\varphi_{m;\sigma}(x)$.

Значение плотности распределения

$$\varphi_{m;\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

случайной величины $X \sim \mathrm{N}(m; \sigma)$ с параметрами m и σ определяется с помощью следующих функций.

Excel: НОРМ.РАСП (x; m; σ; ЛОЖЬ)
MatCalc: nlaw(m,σ).den(x)
R: dnorm (x, m, σ)
Python: gaus = sts.norm(loc = m, scale = σ)
gaus.pdf(x)

Пример. Пусть $X \sim \mathrm{N}(5; 2)$. Значение плотности распределения $\varphi_{5;2}(3) = 0{,}1210$ в точке $x = 3$ в точке находится с помощью следующих функций.

Excel: НОРМ.РАСП (3; 5; 2; 0)
MatCalc: nlaw(5,2).den(3)
R: dnorm(3, 5, 2)
Python: gaus = sts.norm(loc=5, scale=2)
gaus.pdf(3)

χ^2-распределение

1. Процентная точка

100α-процентная точка $\chi^2_{k;\alpha}$ случайной величины, распределенной по закону χ^2 с k степенями свободы, находится с использованием следующих функций.

Excel: ХИ2.ОБР.ПХ(α; k)
MatCalc: x2law(k).q(1-α)
R: qchisq(α, k, lower.tail=FALSE)
Python: chi = sts.chi2(k)
chi.isf(α)

Пример. Пусть $X \sim \chi^2_{20}$ (χ^2 распределение с 20 степенями свободы). 5%-ная точка $\chi^2_{20;\,0,05} = 31,41$ находится с помощью следующих функций.

Excel: ХИ2.ОБР.ПХ(0,05; 20)
MatCalc: x2law(20).q(0.95)
R: qchisq(0.05, 20, lower.tail=FALSE)
Python: chi = sts.chi2(20)
chi.isf(0.05)

2. Квантиль

Квантиль x_α порядка α случайной величины, распределенной по закону χ^2 с k степенями свободы, находится с использованием следующих функций.

Excel: ХИ2.ОБР(α; k)
MatCalc: x2law(k).q(α)
R: qchisq(α, k)
Python: chi = sts.chi2(k)

Пример. Пусть $X \sim \chi^2_{20}$. Квантиль $x_{0,9} = 28,41$ порядка 0,9 находится с помощью следующих функций.

Excel: ХИ2.ОБР(0,9; 20)
MatCalc: x2law(20).q(0.9)
R: qchisq(0.9, 20)
Python: chi = sts.chi2(20)
chi.ppf(0.9)

3. Функция распределения.

Значение функции распределения случайной величины χ^2_k в точке x находится с использованием следующих функций.

Excel: ХИ2.РАСП(x; k; ИСТИНА)
MatCalc: x2law(k).pl(x)
R: pchisq(x,k)
Python: chi = sts.chi2(20)
chi.cdf(x)

Пример. Пусть X имеет χ^2-распределение с 20 степенями свободы. Значение функции распределения $F(31,41) = 0,95$ в точке $x = 31,41$ находится с помощью следующих функций.

Excel: ХИ2.РАСП(31,41;20; 1)
MatCalc: x2law(20).pl(31.41)
R: pchisq(31.41, 20)
Python: chi = sts.chi2(20)
chi.cdf(31.41)

Распределение Стьюдента

Процентная точка.

100α-процентная точка $t_{k;\alpha}$ случайной величины, распределенной по закону Стьюдента с k степенями свободы, находится с использованием следующих функций.

Excel: СТЬЮДЕНТ.ОБР($1 - \alpha$; k)
MatCalk: tlaw(k).q(1-α)
R: qt(α,k,lower.tail=FALSE)
Python: student = sts.t(k)
student.isf(α)

Пример. Пусть $X \sim T_{30}$. Распределение Стьюдента с 30 степенями свободы). 10%-процентная точка $t_{30;0,1}$ находится с помощью следующих функций.

Excel: СТЬЮДЕНТ.ОБР(0,9; 30)
MatCalc: tlaw(30).q(0.9)
R: qt(0.1, 30, lower.tail=FALSE)

Python: student = sts.t(30)

$t_{30;0,1}$ = student.isf(0.1) = 1,3104

2. Квантиль

Квантиль x_α случайной величины, распределенной по закону Стьюдента с k степенями свободы, находится с использованием следующих функций.

Excel: СТЬЮДЕНТ.ОБР(α; k)

MatCalc: tlaw(k).q(α)

R: $t_{k;\alpha}$ = qt(α,k)

Python: student = sts.t(k)

student.ppf(α)

Пример. Пусть $X \sim T_{30}$. Квантиль $x_{0,95} = 1{,}6973$ порядка 0,95 определяется с помощью следующих функций

Excel: СТЬЮДЕНТ.ОБР(0,95; 30) .

MatCalc: tlaw(30).q(0.95)

R: qt(0.95, 30).

Python: student = sts.t(30)

student.ppf(0.95).

3. Функция распределения.

Значение функции распределения $F(x)$ случайной величины T_k в точке x находится с использованием следующих функций.

Excel: СТЬЮДЕНТ.РАСП(x; k; ИСТИНА)

MatCalc: tlaw(k).pl(x)

R: pt(x,k)

Python: student = sts.t(k)

student.cdf(x)

Пример. Пусть X имеет распределение Стьюдента с 30 степенями свободы. Значение функции распределения $F(1{,}3104) = 0{,}9$ в точке $x = 1{,}3104$ находится с использованием следующих функций.

Excel: СТЬЮДЕНТ.РАСП(1,3104; 30; 1)

MatCalc: tlaw(30).pl(1.3104)

R: pt(1.3104, 30)
Python: student = sts.t(30)
student.cdf(1.3104)

F-распределение

1. Процентная точка.

100α-процентную точку $f_{k_1;k_2;\alpha}$ F-распределения с k_1 и k_2 степенями свободы можно найти, вызывая следующие функции.

Excel: F.ОБР.П $X(\alpha; k_1; k_2)$
MatCalc: Flaw(k_1,k_2).q(1−α)
R: qf(α, k_1, k_2,lower.tail=FALSE)
Python: F = sts.f(k_1, k_2)
F.isf(α)

Пример. Пусть $X \sim F_{5;15}$. (F-распределение с 5 и 15 степенями свободы). 5%-процентная точка $f_{5;15;0,05}$ = 2,9013 находится с использованием следующих функций.

Excel: F.ОБР.ПХ(0,05;5;15)
MatCalc: Flaw(5,15).q(0.95).
R: qf(0.05, 5, 15, lower.tail=FALSE)
Python: F = sts.f(5,15)
F.isf(0.05)

2. Квантиль

Квантиль x_α F-распределения с k_1 и k_2 степенями свободы можно найти, вызывая следующие функции.

Excel: F.ОБР.(α; k_1; k_2)
MatCalc: Flaw(k_1,k_2).q(α)
R: qf(α, k_1, k_2).
Python: F = sts.f (k_1, k_2)
F.ppf(α)

Пример. Пусть $X \sim F_{5;15}$. Квантиль $x_{0,9}$ = 2,2730 порядка 0,9 можно найти, используя следующие функции.

Excel: F.ОБР(0,9;5;15).

MatCalc: Flaw(5,15).q(0.9)
R: qf(0.9, 5, 15)
Python: F = sts.f(5,15)
F.ppf(0.9)

3. Функция распределения.

Значение функции распределения $F(x)$ случайной величины $F_{k_1;k_2}$ находится с использованием следующих функций.

Excel: F.РАСП(x; k_1; k_2; ИСТИНА)
MatCalc: Flaw(k_1,k_2).pl(x)
R: pf(x, k_1, k_2)
Python: F = sts.f (x, k_1, k_2)
F.cdf(x)

Пример. Пусть X имеет F-распределение с 5 и 15 степенями свободы. Значение функции распределения $F(2{,}9013) = 0{,}95$ в точке $x = 2{,}9013$ находится с использованием следующих функций.

Excel: F.РАСП(2,9013; 5; 15; 1)
MatCalc: Flaw(5,15).pl(2.9013)
R: pf(2.9013, 5, l5)
Python: F = sts.f(5,15)
F.cdf(2.9013)

7. Примеры

Пример 1.

Случайная величина X имеет показательное распределение с параметром λ.

а) Требуется построить критерий проверки гипотезы $H_0 : \lambda = \lambda_0$ против альтернативы $H_1 : \lambda \neq \lambda_0$, где λ_0 — некоторое заданное число.

Для построения критерия использовать следующее утверждение.

Случайная величина $T = 2\lambda n\overline{X}$, где выборочное среднее $\overline{X}$ вычислено по выборке объема n из показательного распределения с параметром λ, *имеет* χ^2*-распределение с 2n степенями свободы*.

б) Используя построенный критерий, проверить на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : \lambda = 0{,}1$ против альтернативы $H_1 : \lambda \neq 0{,}1$, если по выборке объема $n = 50$ найдено выборочное среднее $\overline{x} = 9{,}5$;

в) построить функцию мощности критерия для значений $\lambda \in [0{,}01;\ 0{,}2]$, и найти мощность для случая, когда альтернатива имеет вид: $H_1 : \lambda = 0{,}15$.

Решение. $H_0 : \lambda = \lambda_0$; $H_1 : \lambda \neq \lambda_0$.

Чтобы построить критерий проверки гипотезы, нужно:

1) задать, по какой формуле будет вычисляться статистика критерия;

2) задать критическую область;

3) убедиться, что для заданной статистики и критической области вероятность ошибки первого рода совпадает с заданным уровнем значимости α.

1. Вид статистики определен в условии задачи.

Следует взять статистику: $T = 2\lambda_0 n\overline{X}$.

В случае справедливости гипотезы $H_0 : \lambda = \lambda_0$ эта статистика распределена по закону: $T \sim \chi^2_{2n}$.

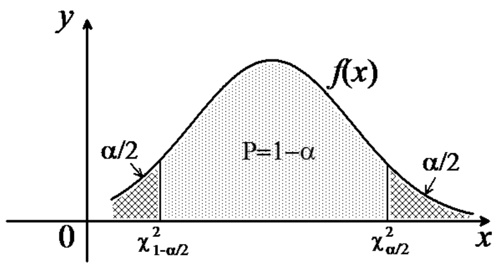
2. Задаем критическую область.

В случае, когда альтернативная гипотеза имеет вид: $H_1 : \theta = \theta_0$, критическую область K следует выбирать двустороннюю.

В критическую область должны входить такие значения статистики, которые расположены слева и справа от области $D = \overline{K}$ принятия гипотезы H_0.

Сама область D должна быть такова, чтобы значение статистики попадало в нее с доверительной вероятностью $1 - \alpha$.

Тогда слева от области D и справа от нее значения статистики должны попадать с вероятностью $\dfrac{\alpha}{2}$.



Это означает, что граничными точками критической области должны быть точки:

$\chi^2_{2n;1-\alpha/2}$ — $100\left(1-\dfrac{\alpha}{2}\right)$-процентная точка; левее этой точки статистика попадает с вероятностью $\dfrac{\alpha}{2}$;

и $\chi^2_{2n;\alpha/2}$ — $100\dfrac{\alpha}{2}$-процентная точка; правее этой точки статистика попадает с вероятностью $\dfrac{\alpha}{2}$.

Итак,

$$K = \left(-\infty; \chi^2_{2n;1-\alpha/2}\right) \cup \left(\chi^2_{2n;\alpha/2}; \infty\right).$$

3. Убедимся, что для выбранной критической области вероятность ошибки первого рода совпадает с α.

Вероятность ошибки первого рода — это вероятность попасть в критическую область, когда справедлива гипотеза $H_0 : \lambda = \lambda_0$, т.е. когда статистика T распределена по закону χ^2_{2n}.

$$P_{H_0}\left(T \in K\right) = P_{H_0}\left(T < \chi^2_{2n;1-\alpha/2}\right) + P_{H_0}\left(T > \chi^2_{2n;\alpha/2}\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

б) Проверяем гипотезу: $H_0 : \lambda = 0{,}1$ против альтернативы: $H_1 : \lambda \neq 0{,}1$.

Чтобы проверить гипотезу, нужно вычислить наблюдаемое значение статистики и проверить, попало ли оно в критическую область. Имеем:

$$T_{\text{набл}} = 2\lambda_0 n\overline{x} = 2 \cdot 0{,}1 \cdot 50 \cdot 9{,}5 = 95.$$

Критические точки:

1) $\chi^2_{2n;1-\alpha/2} = \chi^2_{100;0,975} =$ ХИ2.ОБР.ПХ$\left(0{,}975;100\right) = 74{,}2$;

в R: qchisq(0.975, 100, lower.tail=FALSE) = 74,2;

в MatCalc: x2law(100).q(0.025) = 74,2;

в Python: import scipy.stats as sts
  chi = sts.shi2(100)
  chi.isf(0.975) = 74,2.

2) $\chi^2_{2n;\alpha/2} = \chi^2_{100;0,025} =$ ХИ2.ОБР.ПХ$\left(0{,}025;100\right) = 129{,}6$;

в R: qchisq(0.025, 100) = 129,6;

в MatCalc: x2law(100).q(0.975) = 129,6;

в Python:
  chi = sts.shi2(100)
  chi.isf(0.025) = 74,2.

Итак, $T_{\text{набл}} = 95$; $\chi^2_{2n;1-\alpha/2} = 74{,}2$; $\chi^2_{2n;\alpha/2} = 129{,}6$.

Так как наблюдаемое значение статистики не попало в критическую область: $\chi^2_{2n;\,1-\alpha/2} < T_{\text{набл}} < \chi^2_{2n;\,\alpha/2}$, то гипотеза H_0 принимается.

в) Строим функцию мощности $w(\lambda)$, $\lambda > 0$.

При каждом конкретном значении параметра λ, $w(\lambda)$ — это мощность критерия, когда альтернативная гипотеза имеет вид:

H_1 : параметр распределения Пуассона равен λ.

Мощность критерия — это вероятность принять альтернативную гипотезу, когда эта гипотеза действительно верна, $w = 1 - \beta$.

Вычисляем вероятность $P_{H_1}(T \in K)$.

Статистика T имеет вид: $2\lambda_0 n\overline{X}$.

Когда справедлива гипотеза H_1, распределение случайной величины T нам неизвестно, но известно, что случайная величина $2\lambda n\overline{X}$ имеет распределение χ^2_{2n}.

Для вычисления вероятности $P_{H_1}(T \in K)$, преобразуем статистику.

$$P_{H_1}(T \in K) = 1 - P_{H_1}\left(74,2 < T < 129,6\right) =$$

$$= 1 - P_{H_1}\left(74,2 < 2\lambda_0 n\overline{X} < 129,6\right) =$$

$$= 1 - P_{H_1}\left(74,2 < 2 \cdot 0,1 \cdot n\overline{X} < 129,6\right).$$

С целью получения выражения $2\lambda n\overline{X}$ (с известным законом распределения), умножим все составляющие неравенств на λ:

$$= 1 - P_{H_1}\left(74,2 \cdot \lambda < 2\lambda \cdot 0,1 \cdot n\overline{X} < 129,6 \cdot \lambda\right) =$$

$$= 1 - P_{H_1}\left(74,2 \cdot \frac{\lambda}{0,1} < 2\lambda n\overline{X} < 129,6 \cdot \frac{\lambda}{0,1}\right) =$$

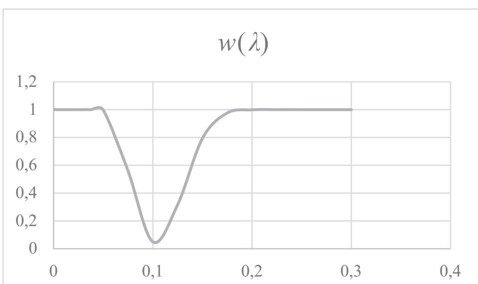
$$= 1 - P\left(742\lambda < \chi^2_{2n} < 1296\lambda\right).$$

Вероятность того, что случайная величина χ^2_{2n} принадлежит заданному интервалу, вычисляем, используя ее функцию распределения.

В Excel — это функция: ХИ2.РАСП(x; v; 1).

$P_{H_1}(T \in K) = 1 - P(742\lambda < \chi^2_{2n} < 1296\lambda) = 1 - (F(1296\lambda) - F(742\lambda) =$
$=$ 1 – (ХИ2.РАСП(1296λ; 100; 1) – ХИ2.РАСП(742λ; 100; 1)).

График полученной функции:



Значение в точке λ = 0,15: w(0,15) = 0,8.

Почему минимальное значение мощности достигается при $\lambda = 0{,}1$, а чем дальше от этого значения, тем ближе мощность к 1?

Сложно различать гипотезы $H_0 : \lambda = \lambda_0$ и $H_1 : \lambda = \lambda_1$, когда параметры λ_0 и λ_1 близки друг к другу.

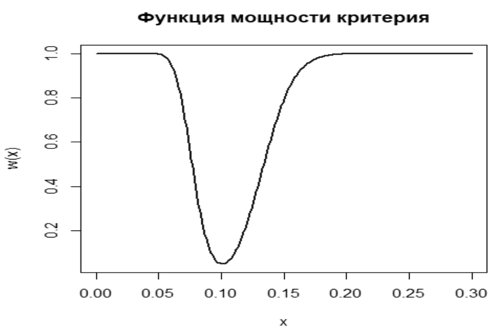
И не сложно, если они далеки друг от друга.

<u>На языке R</u>. Строим график функции

$$w(x) = 1 - (F(1296x) - F(742x))$$

на интервале [0; 0,3].

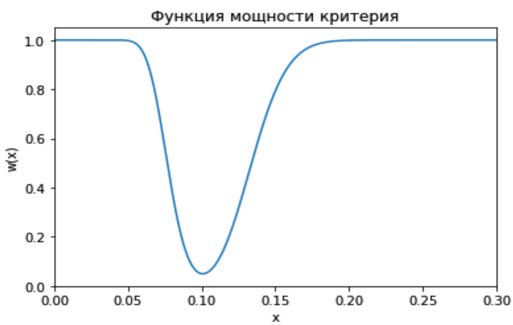
```
x <- seq(0, 0.3, 0.001) # Значения независимой переменной
  # с шагом 0,001.
plot(x, 1-(pchisq(1296*x, 100) – pchisq(742*x, 100)), type="l",
    col="blue", lty=1, pch=2, lwd=2,
   main="Функция мощности критерия", xlab="x", ylab="w(x)")
```



<u>На языке Python</u>.

```
import scipy.stats as sts
%matplotlib inline
import matplotlib.pyplot as plt  # импорт графической библиотеки
import numpy as np  # импорт библиотеки numpy
                    # для работы с массивами
chi = sts.chi2(100)
chi.isf(0.975)
x = np.linspace(0, 0.3, 1000)   # Формирование 1000 значений
# независимой переменной на интервале (0; 0,3)
```

```
plt.plot(x, 1-(chi.cdf(1296*x)-chi.cdf(742*x))) # Построение графика
            # функции w(x)
plt.xlim(0, 0.3) # Задание пределов изменения переменных
plt.ylim(0, 1)
plt.xlabel("x") # Вывод подписей по осям
plt.ylabel("w(x)")
plt.title(u"Функция мощности критерия")) # Вывод название графика
                # Символ "u" позволяет выводить текст на кириллице
```



<u>На языке MatCalc</u>.

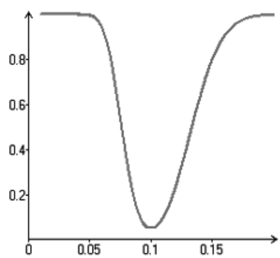
Вычисления в пунктах б) и в) могут быть выполнены с помощью программы:

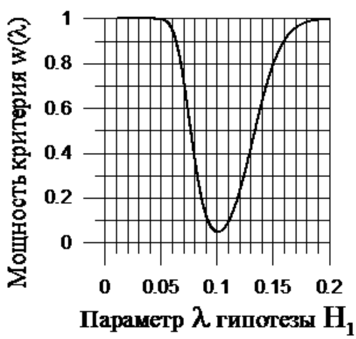
```
//Пример 1.    всё, что следует после //, является комментарием
// Случайная величина X имеет показательное распределение
// с параметром lambda.
// Пусть n — объём выборки, xAv — выборочное среднее.
// б) Проверить гипотезу H0: lambda=L0 против H1:lambda<>L0,
// с уровнем значимости alp,
// опираясь на статистику T=2*n*lambda*xAv ~ xi^2(2n)
alp=0.05; n=50; xAv=9.5; L0=0.1; //исходные величины
x1=x2law(2*n).q(alp/2);   x2=x2law(2*n).q(1-alp/2);   //критические значения
T=2*n*L0*xAv;  // наблюдаемое значение статистики
Str1="D=[x1,x2]=["+x1+", "+x2+"]"; Str2="Tнабл="+T;
```

```
if (T>x1 & T<x2) StrR=Str2+" in "+Str1+" => H0 принимается";
else StrR=Str2+" not in "+Str1+" => H0 отвергается";
message(StrR); // результат проверки гипотезы

// в) Построить функцию мощности критерия
// для значений lamda in [aL; bL]
// и найти мощность для случая, когда H1: lambda=L1
L1=0.15; aL=0.01; bL=0.2;  //исходные величины

function    power(L)=1-(x2law(2*n).pl(L*x2/L0)-x2law(2*n).pl(L*x1/
L0));
w1=power(L1); message("P("+L1+")="+w1);
w0=power(L0); message("P("+L0+")="+w0+"=alp");
L=aL:bL:0.0001; w=power(L);
savetable([L,w],"Power(lambda).txt","eng"); // сохранение результатов
line(L,w,red,2); axes();show();erase(); // построение графика
```





Первый рисунок появляется непосредственно в процессе расчёта, второй построен на основе сохранённых данных специальной программой Grapher.

Пример 2. Имеется монета с вероятностью выпадения орла p.

Проверяется гипотеза $H_0: p = 0{,}5$, альтернативная гипотеза $H_1: p \neq 0{,}5$.

Проверка гипотезы осуществляется по следующей схеме: монета подбрасывается 5 раз. Если орел выпадает 2 или 3 раза, гипотеза H_0 принимается, в противном случае отклоняется.

Требуется:

1) найти величину a — вероятности ошибки первого рода;

2) построить график функции мощности $w(p)$;

3) найти значение мощности критерия при $p = 0{,}3$ и $p = 0{,}55$.

Решение.

В этой задаче статистика критерия: $T = k$, k — число выпадений орла при 5 подбрасываниях монеты.

Критическая область: $k = \{0, 1, 4, 5]$.

а)

$$\alpha = P_{H_0}(H_0 \text{ отклоняется}) = P_{p=0,5}(k \neq 2, k \neq 3) =$$

$$= 1 - \left(P_{p=0,5}(k = 2) + P_{p=0,5}(k = 3)\right) =$$

$$= 1 - \left(C_5^2 \cdot 0{,}5^2 \cdot 0{,}5^3 + C_5^2 \cdot 0{,}5^3 \cdot 0{,}5^2\right) =$$

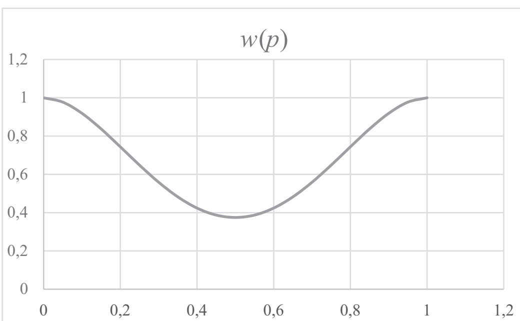
$$= 1 - \frac{2 \cdot 10}{2^5} = \frac{3}{8} = 0{,}375.$$

б)

$$w(p) = P_{H_1}(H_0 \text{ отклоняется}) = P_p(k \neq 2, k \neq 3) =$$

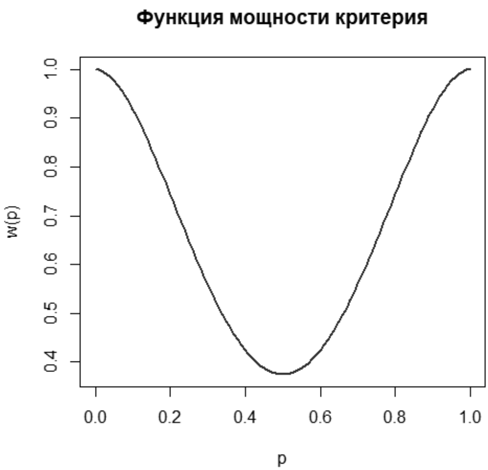
$$= 1 - \left(P_p(k = 2) + P_p(k = 3)\right) =$$

$$= 1 - \left(C_5^2 \cdot p^2 \cdot (1-p)^3 + C_5^2 \cdot p^3 \cdot (1-p)^2\right) = 1 - 10p^2(1-p)^2.$$



На языке R.

```
x <- seq(0, 1, 0.01)  # Значения независимой переменной
 # с шагом 0,01.
plot(x, 1-10*x^2*(1-x)^2, type="l",
    col="blue", lty=1, pch=2, lwd=2,
   main="Функция мощности критерия", xlab="p", ylab="w(p)")
```



На языке Python.

```
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
x = np.linspace(0, 1, 1000)
plt.plot(x, 1-10*x**2*(1-x)**2)
plt.xlim(0, 1)
plt.ylim(0, 1)
plt.xlabel("p")
plt.ylabel("w(p)")
plt.title(u"Функция мощности критерия")
```

Пример 3. Инвестиционный фонд объявил, что средний годовой доход по акциям предприятий некоторой отрасли промышленности составил m_0 = 11,5%. Инвестор, желая проверить, является ли это заявление правильным, осуществил случайную выборку из n = 41 акции этой отрасли. Средний годовой доход по ним составил 10,8%, а исправленное выборочное стандартное отклонение — 3,4%.

(а) Имеет ли инвестор основания, чтобы опровергнуть заявление инвестиционного фонда на 5%-ном уровне значимости.

(б) Найдите P — значение критерия.

Для проверки гипотезы против гипотезы $H_1: m < m_0$ используется статистика с левосторонней критической областью , где $T_{кр} = -t_{n-1;\,\alpha}$ ($t_{n-1;\,\alpha}$— 100α-процентная точка распределения Стьюдента с $n - 1$ степенью свободы).

При решении задачи используйте тот факт, что при условии справедливости гипотезы H_0 статистика T имеет закон распределения Стьюдента с $n - 1$ степенями свободы.

Решение. $H_0: m = 11{,}5$; $H_1: m < 11{,}5$; $n = 41$; = 10,8%; $s = 3{,}4\%$.

а)

;

$T_{кр} = -t_{n-1;\,\alpha} = -t_{40;\,0,05}$.

В Excel: $-t_{40;\,0,05}$ = – СТЬЮДЕНТ.ОБР(0,95; 40) = –1,68

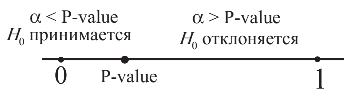
В MatCalc: $-t_{40;\,0,05}$ = tlaw(40).q(0.95) = –1,68

На языке R: $-t_{40;\,0,05}$ = –qt(0.05,40,lower.tail=FALSE) = –1,68
На языке Python: student = sts.t(40)
–student.isf(0.05) = –1,68

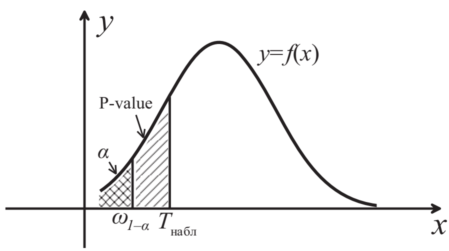
Так как $T_{набл} = T_{кр}$, то гипотеза H_0 принимается; то есть оснований для опровержения заявления нет.

б) Ищем P-значение критерия.

(P-значение критерия: минимальная граница значений уровня значимости α, при которых нулевая гипотеза отклоняется).



Для левосторонней критической области P-значение вычисляется по формуле: $P - value = P_{H_0}(T < T_{набл})$.



При условии справедливости гипотезы H_0, статистика T имеет распределение Стьюдента с $n - 1 = 40$ степенями свободы.

Искомую вероятность можно получить

в Excel: СТЬЮД.РАСП($T_{набл}$; $n-1$; 1) = СТЬЮД.РАСП(–1,32; 40; 1) = 0,097;

в MatCalc: tlaw(40).pl(–1.32)

на языке R: pt(–1.32, 40) = 0.097

на языке Python: student = sts.t(40)

student.cdf(–1.32) = 0,097

Так как значение уровня значимости $\alpha = 0{,}05 < P - value$, то для этого значения α гипотеза H_0 принимается.

Ответ: а) нет, не имеет; б) ≈ 0,097.

Пример 4. Завод рассылает рекламные каталоги возможным заказчикам.

Как показал опыт, вероятность p того, что организация, получившая каталог, закажет рекламируемое изделие, равна величине p_0, где $p_0 = 0{,}08$. Завод разослал n = 1000 каталогов новой усовершенствованной формы и получил k = 100 заказов.

(а) Можно ли на уровне значимости $a = 0{,}05$ считать, что новая форма рекламы оказалась эффективнее первой?

(б) Найдите P – значение критерия.

Для проверки гипотезы $H_0 : p = p_0$ против гипотезы $H_1 : p > p_0$ используется статистика $Z = \dfrac{\frac{k}{n} - p_0}{\sqrt{p_0(1-p_0)/n}}$ с правосторонней критической областью $K = \left\{\vec{x} : Z(\vec{x}) > Z_{\text{кр}}\right\}$,

где $Z_{\text{кр}} = z_{\alpha;}$ (z_α — 100α-процентная точка стандартного нормального распределения).

При больших значениях параметра n, закон распределения статистики Z имеет приближенно стандартное нормальное распределение $N(0; 1)$.

Решение.

Пусть p — вероятность заказа изделия.

Исходные данные: $H_0 : p = 0{,}08$; $H_1 : p > 0{,}08$; $n = 1000$;

$$\frac{k}{n} = \frac{100}{1000} = 0{,}1\,;\ p_0 = 0{,}08;\ a = 0{,}05.$$

Проверяется гипотеза $H_0 : p = 0{,}08$, означающая, что новая форма рекламы не лучше старой.

а) Находим наблюдаемое значение статистики — значение статистики критерия на данных выборки.

$$Z_{\text{набл}} = \frac{\frac{k}{n} - p_0}{\sqrt{p_0\left(1-p_0\right)/n}} = \frac{0,1-0,08}{\sqrt{0,08\left(1-0,08\right)/1000}} \approx 2,33.$$

Находим критическое значение статистики — это 5%-ная точка $z_{0,05}$ нормального закона распределения:

$Z_{\text{кр}} = z_{\alpha} = z_{0,05} = 1,64.$

В Excel: $z_{0,05}$ = НОРМ.СТ.ОБР(0,95) = 1,64

В MatCalc: nlaw(0,1).q(0.95) = 1,64

На языке R: qnorm(0.05, lower.tail=FALSE) = 1,64

На языке Python: sts.norm.isf(0.05) = 1,64

Итак, имеем значения: $Z_{\text{набл}} = 2,33$; $Z_{\text{кр}} = 1,64$

Будет принято решение об отклонении нулевой гипотезы, так как наблюдаемое значение статистики больше критического (наблюдаемое значение попало в критическую область).

б) Найдем P-значение критерия.

Для правосторонней критической области P-значение вычисляется по формуле:

$P - value = P_{H_0}(Z > Z_{\text{набл}}).$

При условии справедливости гипотезы H_0, статистика Z имеет стандартное нормальное распределение.

Искомую вероятность можно получить, используя функцию Лапласа

$P_{H_0}(Z > Z_{\text{набл}}) = P(Z > 1,33) = 0,5 - \Phi\left(1,33\right) = 0,092$

В Excel: 1 – НОРМ.СТ.РАСП(1,33; 1) = 0,092

В MatCalc: 1 – nlaw(0,1).pl(1.33) = 0,092

На языке R: 1 – pnorm(1.33) = 0,092

На языке Python: 1 – sts.norm.sdf(1.33) = 0,092

Так как выполняется неравенство: $a = 0,05 > P - value$, то для заданного значения уровня значимости гипотеза H_0 отклоняется.

Ответ: а) да, можно; б) 0,092.

Пример 5. В десятичной записи числа π среди первых 10002 знаков после запятой цифры 0, 1, …, 9 встречаются соответственно 968, 1026, 1021, 974, 1014, 1046, 1021, 970, 948, 1014 раз.

Проверьте на уровне значимости $a = 0{,}05$, согласуются ли эти данные с гипотезой H_0 о равномерном распределении цифр.

Решение.

Используем критерий χ^2.

Объем выборки $n = 10002$; число интервалов: $k = 10$; все значения теоретических вероятностей $p_i = \dfrac{1}{10}$.

Наблюдаемое значение статистики находится по формуле

$$\chi^2_{\text{набл}} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

где n_i – наблюдаемые частоты;

p_i – теоретические вероятности;

k – число интервалов (число различных значений в выборке для дискретной случайной величины).

Критическое значение статистики вычисляется по формуле

$$\chi^2_{\text{кр}} = \chi^2_{к-l-1,\,\alpha},$$

где $\chi^2_{k-l-1,\,\alpha}$ -100a процентная точка χ^2-распределения с $k - l - 1$ степенями свободы порядка α;

χ^2 – уровень значимости;

l – количество параметров распределения, оцениваемых по выборке (в данном примере таких параметров нет, $l = 0$).

Если наблюдаемое значение статистики, вычисляемое по выборке, оказывается больше критического, $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$, то гипотеза H_0 отклоняется на уровне значимости α.

<u>Выполняем вычисления *в Excel*</u>. Исходные данные, и промежуточные величины, необходимые для нахождения наблюдаемого значения статистики, удобно помещать в таблицу следующего вида:

I	n_i	p_i	$n_i p_i$	$n_i - np_i$	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
1	968	0,1	1000,2	–32,2	1036,84	1,03663
2	1026	0,1	1000,2	25,8	665,64	0,66551
3	1021	0,1	1000,2	20,8	432,103	0,43255
4	974	0,1	1000,2	–26,2	686,44	0,68630
5	1014	0,1	1000,2	13,8	190,44	0,19040
6	1046	0,1	1000,2	45,8	2097,64	2,09722
7	1021	0,1	1000,2	20,8	432,64	0,43255
8	970	0,1	1000,2	–30,2	912,04	0,91186
9	948	0,1	1000,2	–52,2	2724,84	2,72430
10	1014	0,1	1000,2	13,8	190,44	0,19040
Σ	10002	1	10002	0	–	9,36773

В последней строке таблицы вычисляется сумма значений по соответствующему столбцу. Такая сумма значений строк последнего столбца дает наблюдаемое значение статистики:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{10} \frac{(n_i - np_i)^2}{np_i} = \frac{\left(968 - 10002 \cdot \frac{1}{10}\right)^2}{10002 \cdot \frac{1}{10}} + \ldots + \frac{\left(1014 - 10002 \cdot \frac{1}{10}\right)^2}{10002 \cdot \frac{1}{10}} = 9{,}36773.$$

Критическое значение статистики:

$$\chi^2_{\text{кр}} = \chi^2_{r-1;\,\alpha} = \chi^2_{9;\,0,05} = \text{ХИ2.ОБР.ПХ}(0{,}05;\ 9) = 16{,}92.$$

Наблюдаемое значение меньше критического. Нулевая гипотеза на 5 %-м уровне принимается.

Найдем P-значение. В данном примере критическая область статистики — правосторонняя, и P-значение вычисляется как вероятность того, что случайная величина, распределенная по закону χ^2 с 9 степенями свободы, примет значение больше, чем наблюдаемое значение.

P-value = $P(\chi_k^2 > \chi^2_{\text{набл}}) = P(\chi_9^2 > 9{,}36773)$ = ХИ2.РАСП.ПХ (514,94; 5) = 0,404.

Нулевая гипотеза будет приниматься на любом уровне значимости, не превосходящем 0,404, в частности при $a = 0{,}05$.

Решение задачи на языке Python.

Проверку гипотезы по критерию χ^2 в языке Python осуществляет функция chisquare модуля scipy.stats.

Импортируем библиотеки: scipy.stats — для вызова программы расчета по критерию с χ^2 (sts.chisquare); numpy – для работы с массивами.

```
import scipy.stats as sts
import numpy as np
n = 10002 # Объем выборки
observed_frequences = np.array([968, 1026, 1021, 974, 1014,
1046, 1021, 970, 948, 1014]) # Формирование массива исходных
                                # данных (значения ni)
M = np.full(10, 0.1)  # массив длиной 10 с теоретическими вероятностями pi = 0,1
expected_frequences = M*n  # ожидаемые частоты n*pi
chisq, p = sts.chisquare(f_obs = observed_frequences,
                         f_exp = expected_frequences)
 # параметры: f_obs – наблюдаемые частоты, f_exp – теоретические частоты
 # Выходные параметры функции:
 # первый параметр, chisq – наблюдаемое значение статистики,
 # второй параметр p – P-значение
print ('chi2 = ', chisq)
print ('p-value =', p)
```

Результат работы программы:

chi2 = 9.36772645471

p-value = 0.404045207515

Наблюдаемое значение статистики: 9.37, P-значение: 0.4. Так как P-значение больше 0.05, гипотеза H_0 принимается.

Ответ: Гипотеза H_0 принимается.

Пример 6. Сгенерируйте две независимые выборки: $Y \sim \text{Exp}(1)$, $Z \sim \text{Exp}(1)$ распределенные по показательному закону с параметром $\lambda = 1$.

Объёмы выборок — по 200 элементов.

Сформируйте новую выборку X по правилу: $X = Y + Z$.

Проверьте, используя подходящие критерии, гипотезу $H_0 : X \sim \text{Exp}(0{,}5)$ о соответствии выборки X показательному закону распределения с параметром $\lambda = 0{,}5$.

(Принять уровень значимости $\alpha = 0{,}05$).

Решение. Используем критерий Колмогорова (одновыборочный критерий Колмогорова-Смирнова).

Решение в Excel.

	A	B	C	*D*	*E*	*F*	*G*	*H*	*I*	J	*K*	L
1	*i*	u1	u2	*y*	*z*	*x*	$x_{(i)}$	*F* (*x*)	*i/n*	*(i-1) /n*	*i/n — F (x)*	*F* (*x*) — (*i*–1) /*n*
2	0											
3	1	0,075	0,774	0,078	1,485	1,564	0,089	0,043	0,005	0,000	-0,038	0,043
4	2	0,491	0,740	0,676	1,346	2,022	0,110	0,053	0,010	0,005	-0,043	0,048
5	3	0,622	0,412	0,972	0,531	1,503	0,179	0,085	0,015	0,010	-0,070	0,075
6	4	0,614	0,159	0,951	0,174	1,125	0,213	0,101	0,020	0,015	-0,081	0,086
7	5	0,951	0,970	3,009	3,498	6,506	0,309	0,143	0,025	0,020	-0,118	0,123
8	6	0,856	0,474	1,939	0,643	2,582	0,330	0,152	0,030	0,025	-0,122	0,127
9	7	0,119	0,747	0,126	1,375	1,501	0,354	0,162	0,035	0,030	-0,127	0,132
	.	…	…	…	…	…	…	…	…	…	…	…

1. Формирование массива X.

Генерируем последовательности: $u_{1i} \sim \text{Unif}(0; 1)$ (столбец B, команда СЛЧИС()), $u_{2i} \sim \text{Unif}(0; 1)$ (столбец C, команда СЛЧИС()).

Для генерирования экспоненциально распределенных величин используем следующий факт: если $u \sim \text{Unif}(0; 1)$, то $-\dfrac{1}{\lambda}\ln(1-u) \sim \text{Exp}(\lambda)$.

Формируем последовательности:
$y_i \sim \mathrm{Exp}(1)$ (столбец D, в ячейке D3 команда: =-LN (1-B3) и
$z_i \sim \mathrm{Exp}(1)$ (столбец D, в ячейке E3 команда: =-LN (1-C3).
Формируем массив $x_i = y_i + z_i$ (столбец F).

2. Проверяемая гипотеза: $H_0 : X \sim \mathrm{Exp}(0{,}5)$. Альтернативная гипотеза H_1 — отрицание гипотезы H_0.

В критерии Колмогорова рассматривается статистика следующего вида:

$$D_n = \sqrt{n} \cdot \sup_x \left| F^*(x) - F(x) \right|,$$

где $F(x)$ — гипотетическая функция распределения (в нашем примере — функция распределения экспоненциальной случайной величины с параметром $\lambda = 0{,}5$; это функция: $F(x) = 1 - e^{-0{,}5x}$);

$F^*(x)$ — эмпирическая функция распределения, построенная по выборочным данным.

Критические точки статистики критерия Колмогорова $D_{\text{кр}}$ для некоторых значений уровня значимости α приведены в таблице:

α	0,15	0,1	0,05	0,025	0,01	0,005	0,001
$D_{\text{кр}}$	1,138	1,224	1,358	1,480	1,628	1,731	1,950

Гипотеза H_0 отклоняется на уровне значимости α, если вычисленное по выборке $x_1, ..., x_n$ значение статистики D_n удовлетворяет неравенству $D_n > D_{\text{кр}}$.

Вычисление наблюдаемого значения статистики можно проводить следующим образом.

Предварительно находятся значения

$$D_n^+ = \max_{1 \le i \le n} \left\{ \frac{i}{n} - F_0\left(x_{(i)}\right) \right\}; \; D_n^- = \max_{1 \le i \le n} \left\{ F_0\left(x_{(i)}\right) - \frac{i-1}{n} \right\}.$$

Затем вычисляется D_n:

$$D_n = \sqrt{n} \cdot \max\left\{ D_n^+, D_n^- \right\}.$$

Приступаем к реализации указанного алгоритма.

Формируем вариационный ряд: $x_{(i)}$ (В столбце G — отсортированный массив X).

Формируем последовательность: $F\left(x_i\right)=1-e^{-0,5x_{(i)}}$ (столбец H, в H3 — команда =1-EXP (–0,5*G3).

В столбец I помещаем значения $i/n(n=200)$ (в ячейке I3 команда =A3/200).

В столбец J помещаем значения (i–1)/n (в ячейке J3 команда =A2/200).

Формируем массив $\frac{i}{n}-F(x_{(i)})$ (столбец K, в ячейке K3 команда: =I3–H3).

Формируем массив $F(x_{(i)})-\frac{i-1}{n}$ (столбец L, в ячейке L3 команда: =H3–J3).

Вычисляем величину $D_n^+=\max\limits_i\left(\frac{i}{n}-F(x_{(i)})\right)$ (по команде =МАКС (K3:K202).

Вычисляем величину $D_n^-=\max\limits_i\left(F(x_{(i)})-\frac{i-1}{n}\right)$ (по команде =МАКС (L3:L202).

Результат:

$$D_n^+ = 0{,}0739;\ D_n^- = 01710.$$

Наблюдаемое значение статистики D_n вычисляем по формуле

$$D_n = \max\{D_n^+, D_n^-\}\sqrt{n}.$$

Результат: $D_n = 2{,}4182$.

Для 5%-го уровня значимости $D_{кр} = 1{,}358$.

Наблюдаемое значение статистики D_n больше критического $D_{кр}$. Гипотеза H_0 отклоняется.

Решение на языке Python.

```
import numpy as np
import scipy.stats as sts # загрузка необходимых модулей
exp_rv = sts.expon (scale=1) # задание экспоненциально
            # распределенной случайной величины с параметром 1
np.random.seed (10) # для воспроизводимости результатов
n = 200
Y = exp_rv.rvs (n) # генерирование выборки объема 200
                    # с экспоненциальным распределением Exp (1)
```

```
Z = exp_rv.rvs (n)
X1 = Y + Z
X = np.sort (X1)
FX = 1 – np.exp(–0.5*X)

In1 = np.arange(1, n+1)/n

In2 = np.arange(0,n)/n

Dn1 = np.max(In1–FX)

print(Dn1)
  # Результат: 0.060237971290968861

Dn2 = np.max(FX-In2)
print(Dn2)
  # Результат:  0.1836021368099004

Dn = max(Dn1, Dn2)*np.sqrt(n)  # наблюдаемое значение статистики Dn
print(Dn)
Результат: 2.0357309060655249
```

В языке Python в модуле scipy.stats имеется функция kstest(), реализующая проверку гипотезы H_0 с помощью критерия Колмогорова-Смирнова.

Соответствующая команда имеет вид:

```
sts.kstest(X, 'expon', [0, 2])
  # Первый аргумент – массив выборочных данных.
  # Второй аргумент – строка, указывающая тип
        # гипотетического распределения.
   # Третий аргумент – параметры гипотетического распределения;
  # должны быть оформлены в виде списка (в языке Python –
  # в квадратных скобках.
```

```
# Для экспоненциального распределения первым параметром
# является параметр сдвига (у нас: 0), вторым – параметр
# масштаба, 1/λ (у нас: 1/0,5=2)
# В качестве выходных данных функция дает:
#  1) одно из значений Dn+ , Dn– или Dn;  2) P-значение.
```

Результат:
KstestResult(statistic=0.1836021368099,
pvalue=2.2558671199934e-06)

Решение на языке R.

```
n = 200
set.seed(99) # для воспроизводимости результатов
Y <– rexp(n=200, rate=1) # генерирование выборки объема 200
 # с экспоненциальным распределением Exp(1). Параметр rate
 # задает значение параметра экспоненциального распределения
 # (rate = 1/E(X))
Z <– rexp(200, 1)
X1 = Y + Z
X = sort(X1)
FX = 1 – exp(–0.5*X)
In1 = seq(1,n)/n
In2 = seq(0,n–1)/n
Dn1 = max(In1 – FX)
print(Dn1)
   # Результат: 0.05295457
Dn2 = max(FX–In2)
print(Dn2)
   # Результат: 0.1222877

Dn = max(Dn1, Dn2)*sqrt(n)  # наблюдаемое значение статистики
print(Dn)
   # Результат: 1.72941
```

Наблюдаемое значение статистики больше критического, гипотеза H_0 отклоняется.

В языке R имеется функция kstest(), реализующая проверку гипотезы H_0 с помощью критерия Колмогорова-Смирнова.

Соответствующая команда имеет вид:

ks.test(X, "pexp", 0.5)   # Первый аргумент — массив выборочных данных.

Второй аргумент — строка, указывающая тип гипотетического распределения,

для экспоненциального распределения имеет вид: "pexp".

Третий аргумент — параметры гипотетического распределения.

Результат вызова функции:

One-sample Kolmogorov-Smirnov test

data: X

D = 0.12229, p-value = 0.005049

alternative hypothesis: two-sided

И здесь нулевая гипотеза отклоняется.

Расчёты на MatCalc можно провести с помощью следующей программы:

//Пример 6.

// Сгенерируйте две независимые выборки

// объёмом по $N = 200$ элементов: $Y \sim$ Exp(1), $Z \sim$ Exp(1),

// распределенные по показательному закону с параметром L=1.

// Сформируйте новую выборку X по правилу: $X = Y + Z$.

// Проверьте, используя подходящие критерии,

// гипотезу H_0 о соответствии выборки X

// показательному закону распределения с параметром L=0.5.

// Принять уровень значимости alp=0.05.

alp=0.05; N=200;

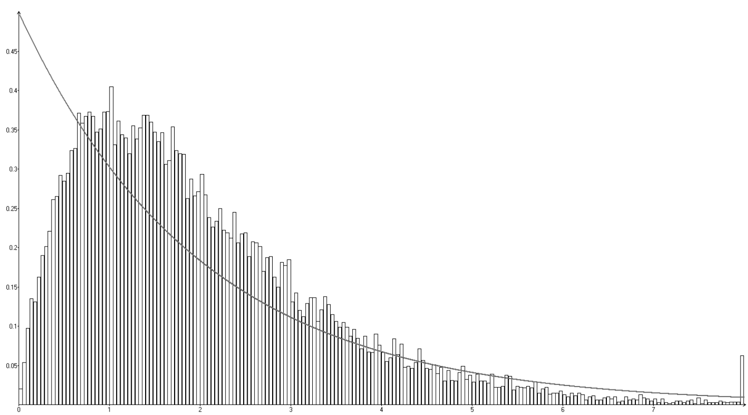
initrand(111222333); //инициализация датчика псевдослучайных чисел

Y=explaw(1)(N); Z=explaw(1)(N); // Две выборки из

// экспоненциального распределения

$X = Y + Z$;

```
xv=sort(X); Fv=1–exp(–0.5*xv); Fe1=(1:N)/N; Fe2=(0:N–1)/N;
Dp=max(Fe1–Fv); Dn=max(Fv–Fe2); D=max([Dp,Dn])*N^0.5;
Dcr=crKolm(alp);
Str1="D+="+Dp+", D–="+Dn+", D="+D;  Str2="Dcr="+Dcr;
if (D < Dcr) StrR=Str1+" < "+Str2+" => H0 принимается";
else StrR=Str1+" > "+Str2+" => H0 отвергается";
message(StrR); // результат проверки гипотезы
    // Дальше только графики:
fe = X.intfr(0:8:0.04);
fe=[fe.c(1:2),fe.c(3)/N]; // группировка плотности логдоходности
    //структура f_denSum1.txt: [x1,x2,f1; x2,x3,f2; x3,x4,f3; ...]
//savetable(fe,"fe_den.txt","eng");
//wintitle("Эмпирическая плотность X");
dhist(fe, white,0.9);    // гистограмма эмпирической плотности
xx=0:8:0.001; ff=L*exp(-L*xx);
line(xx,ff,red,2); axes(); show(); erase();
```



Пример 7. Экзаменационный билет по математике содержит 10 заданий.

Предполагается, что вероятность решения любой отдельной задачи не зависит от исхода решения других задач и что эта вероятность p одна и та же для всех задач. Пусть X — случайная величина, равная числу задач, решенных абитуриентами на вступительном экзамене.

Результаты сдачи экзамена по математике для 300 абитуриентов представлены в таблице.

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	13	17	15	35	10	9	40	51	45	33	32

При этих предположениях проверьте гипотезу H_0 о биномиальном законе распределения случайной величины X на уровне значимости $\alpha = 0{,}05$.

Решение.

Для проверки используем критерий χ^2.

Объем испытаний: $N = 300$. Параметры биномиального закона: $n = 10$ и p.

Если случайная величина X имеет биномиальное распределение с параметрами $(n; p)$, то вероятность того, что абитуриент решит k задач, равна

$P(X = k) = C_n^k p^k q^{n-k}$, где $q = 1 - p$.

Для биномиально распределенной случайной величины X: E $(X) = np$. Поэтому оценкой величины p является $\hat{p} = \dfrac{\overline{x}}{n}$.

Находим $\overline{x} = \dfrac{1}{300}\left(0 \cdot 13 + \ldots + 10 \cdot 32\right) = \dfrac{1811}{300} \approx 6{,}03667$.

Следовательно, $\hat{p} = \dfrac{\overline{x}}{n} = \dfrac{6{,}03667}{10} = 0{,}60367$.

Итак, теоретические вероятности p_i того, что случайная величина X принимает значение x_i вычисляются по формуле

$$p_i = C_{10}^{x_i} \cdot 0{,}60367^{x_i} \cdot 0{,}39633^{10-x_i}.$$

Заполняем таблицу.

i	x_i	n_i	p_i	Np_i	$n_i - Np_i$	$\dfrac{(n_i - Np_i)^2}{Np_i}$
1	0	13	0,0000956	0,02869		
2	1	17	0,001457	0,436982		
3	2	15	0,009984	2,995111		
4	3	35	0,040551	12,16517		
5	4	10	0,108087	32,42597		
6	5	9	0,197556	59,2667		
7	6	40	0,250752	75,22568		
8	7	51	0,218245	65,47337		
9	8	45	0,124655	37,39664		
10	9	33	0,042193	12,65775		
11	10	32	0,006426	1,927939		
Σ	–	300	1	300		

Так как в 1, 2, 3 и 11 группах теоретическая частота $Np_i < 5$, то группы 1, 2 и 3 объединяем с 4-й, а группу 11 объединяем с 10-й. В результате, получим

i	x_i	n_i	p_i	Np_i	$n_i - Np_i$	$\dfrac{(n_i - Np_i)^2}{Np_i}$
1–4	0–3	80	0,05209	15,62596	64,37404	265,20089
5	4	10	0,10809	32,42597	–22,45597	15,50991
6	5	9	0,19756	59,26670	–50,26670	42,63341
7	6	40	0,25075	75,22568	–35,22568	16,49501
8	7	51	0,21824	65,47337	–14,47337	3,19944
9	8	45	0,12466	37,39664	7,60336	1,54589
10–11	9–10	65	0,04862	14,58569	50,41431	174,25314
Σ	–	300	1	300	0	518,84

Наблюдаемое значение статистики:

$$\chi^2_{\text{набл}} = 518{,}84.$$

Критическое значение статистики:

$$\chi^2_{\text{кр}} = \chi^2_{n-r-1;\,\alpha} = \chi^2_{7-1-1;\,0{,}05} = 11{,}07.$$

Так как $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$, то гипотеза H_0 отклоняется.

Найдем P-значение.

$$\text{P-value} = P(\chi^2_{\text{кр}} > \chi^2_{\text{набл}}) = P(\chi^2_5 > 514{,}94) =$$

$$= \text{ХИ2.РАСП.ПХ } (514{,}94;\ 5) = 6{,}8502\text{E-}110.$$

P-значение практически равно 0. Оно меньше значения $\alpha = 0{,}05$, и на этом уровне значимости гипотеза H_0 отклоняется.

<u>Решение примера в Excel</u>.

	A	B	C	D	E	F	G	H	I	J	K
1	i	x_i	n_i	p_i	$N*p_i$	$n_i - Np_i$	$(n_i - Np_i)^2$	$(n_i - Np_i)^2/Np_i$		n =	10
2	1	0	13	9,6E-05	0,02869	12,97131	168,254889	5864,62757		N =	300
3	2	1	17	0,00146	0,436982	16,56302	274,333555	627,790995			
4	3	2	15	0,00998	2,995111	12,00489	144,117354	48,1175294		p =	0,60367
5	4	3	35	0,04055	12,16517	22,83483	521,429327	42,8624673			
6	5	4	10	0,10809	32,42597	–22,42597	502,923979	15,5099147		q =	0,39633
7	6	5	9	0,19756	59,2667	–50,2667	2526,7415	42,6334069			
8	7	6	40	0,25075	75,22568	–35,22568	1240,84854	16,4950125			
9	8	7	51	0,21824	65,47337	–14,47337	209,478328	3,19944337			
10	9	8	45	0,12466	37,39664	7,603364	57,8111495	1,54589172			
11	10	9	33	0,04219	12,65775	20,34225	413,807015	32,6919807			
12	11	10	32	0,00643	1,927939	30,07206	904,328874	469,065170			
13											
□	□		300	1	300	–4,6E-14		**7164,54**			

Вероятности $p_{\text{i}} = P(X = i) = C_n^i p^i q^{n-i}$ биномиального распределения вычисляются с помощью функции БИНОМ.РАСП. В ячейке D2 стоит команда:

БИНОМ.РАСП (B2;K1;K4;1)

Оценка вероятности успеха $\hat{p}$ в ячейке K4 вычисляется по формуле

СУММПРОИЗВ (B2:B12; C2:C12)/(K1*K2)

В ячейке H14 вычисляется наблюдаемое значение статистики, команда СУММ (H2:H12) — сумма значений по строкам столбца H. Получено значение $\chi^2_{\text{набл}} = 7164{,}54$.

Так как в 1, 2, 3 и 11 группах (ячейки E2–E4 и E12) теоретическая частота $np_i < 5$, то группы 1, 2 и 3 объединяем с 4-й, а группу 11 объединяем с 10-й. В результате, получим

A	B	C	D	E	R	G	H
i	x_i	n_i	p_i	$N{*}p_i$	$n_i - Np_i$	$(n_i - Np_i)^2$	$(n_i - Np_i)^2/Np_i$
1–4	0–3	80	0,052	15,626	64,374	4144,018	265,201
5	4	10	0,108	32,426	–22,426	502,924	15,510
6	5	9	0,198	59,267	–50,267	2526,741	42,633
7	6	40	0,251	75,226	–35,226	1240,849	16,495
8	7	51	0,218	65,473	–14,473	209,478	3,199
9	8	45	0,125	37,397	7,603	57,811	1,546
10–11	9–10	65	0,049	14,586	50,414	2541,602	174,253
□		300	1	300	0		**518,84**

P-value = 7E-110.

Сумма значений по столбцу H дает наблюдаемое значение статистики

$$\chi^2_{\text{набл}} = 514{,}94.$$

P-значение вычисляется по команде

P-value =ХИ2.РАСП.ПХ (518,84; 5) = 7E-110.

Решение примера на языке Python.

Импортируем библиотеки: scipy.stats — для вызова программы расчета по критерию χ^2 (sts.chisquare) и расчета вероятностей биномиального распределения (sts.binom); numpy – для работы с массивами.

```
import scipy.stats as sts
import numpy as np
N = 300
n = 10
X = np.arange(0, 11) # значения случайной величины от 0 до 10
observed_frequences = np.array([13, 17, 15, 35, 10, 9, 40, 51, 45,
33, 32])              # наблюдаемые частоты значений
x_v = np.sum(observed_frequences*X)/N # выборочное среднее
p_v = x_v/n
P = np.zeros(11)
binomial_rv = sts.binom(n, p_v) # функция для работы
# с биномиальным распределением
for i in range (11):  # вычисление вероятностей pi
  P[i] = binomial_rv.pmf(i) # функция binomial_rv.pmf вычисляет
                        # веротности pi биномиального распределения
expected_frequences = N*P # ожидаемые частоты N*pi
chisq, p = st.chisquare(f_obs = observed_frequences,
 f_exp = expected_frequences, ddof=1)
# программа расчета по критерию хи-квадрат. Параметр ddof задает
# число параметров распределения, оцениваемых по выборке
# (оценивалась вероятность p)
print('chi2 = ', chisq)
print('p-value =', p)
```

Результат работы программы:

```
chi2 = 7164.53938864
p-value = 0.0
```

Программа sts.chisquare, выполняющая расчет согласно критерию χ^2, не выполняет проверку условия $np_i \geq 5$. Нужно самим проверить и внести соответствующие поправки.

Получив массив expected_frequences, распечатываем его содержимое:

```
expected_frequences
array([ 2.86897824e-02,  4.36982303e-01,  2.99511125e+00,
        1.21651729e+01,  3.24259666e+01,  5.92667037e+01,
        7.52256801e+01,  6.54733661e+01,  3.73966356e+01,
        1.26577529e+01,  1.92793865e+00])
```

Имеем: $np_1 = 12{,}65775$, $np_2 = 0{,}043698$, $np_3 = 2{,}995111$, $np_4 = 12{,}16517$, $np_{10} = 12{,}65775$, $np_{11} = 1{,}92794$. Требуемое неравенство не выполняется для первых трех элементов массива, объединяем их с четвертым; и для предпоследнего элемента, объединяем его с последним.

```
expected_frequences_new = expected_frequences[3:10]
  # в новом массиве ожидаемых частот отсутствуют
  # первые две и последний элементы старого массива
expected_frequences_new[0] = np.sum(expected_frequences[0:4])
  # первым элементом нового массива является
  # сумма первых четырех элементов старого массива
expected_frequences_new[6] = np.sum(expected_frequences[9:11])
  # последним элементом нового массива является
  # сумма двух последних элементов старого массива
```

Такие же действия совершаем с массивом observed_frequences:

```
observed_frequences_new = observed_frequences[3:10]
observed_frequences_new[0] = np.sum(observed_frequences[0:4])
observed_frequences_new[6] = np.sum(observed_frequences[9:11])
```

Вызываем функцию st.chisquare:

```
chisq, p = sts.chisquare(f_obs = observed_frequences_new, f_exp = expected_frequences_new, ddof=1)
print('chi2 = ', chisq)
print('p-value =', p)
```

Результат:
chi2 = 518.837694418
p-value = 6.85016404999e-110

Гипотеза H_0 отклоняется.

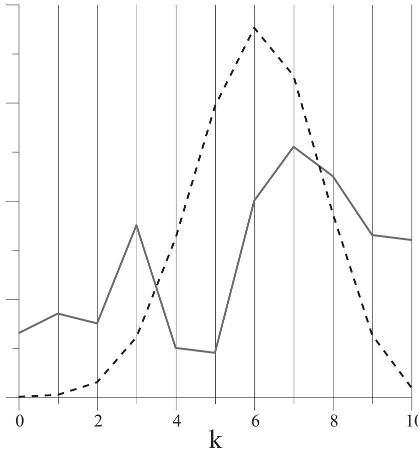
В разобранной выше схеме проблема использования интервалов, которые предположительно должны содержать слишком малое (менее 5) количество событий, решалась путём присоединения таких интервалов к соседним, более «населённым».

На самом деле это может привести к снижению роли относительно маловероятных событий.

Посмотрим, к чему приведёт другой подход, при котором в один интервал объединяются все «бедные» интервалы.

i	x_i	n_i	pi	$n*p_i$	x_i	n_i	pi	$n*p_i$	n_i-$n*p_i$	(ni-npi)^2/npi	x_i	n_i	pi	$n*p_i$	n_i-$n*p_i$	(ni-npi)^2/npi
1	0	13	0,0001	***0,029***												
2	1	17	0,0015	***0,437***												
3	2	15	0,0100	***2,995***												
4	3	35	0,0406	12,165							0-2,10	77	0,018	5,38872	71,61	951,6
					0-3	80	0,0521	15,626	64,374	265,201	3	35	0,0406	12,1652	22,83	42,86
5	4	10	0,1081	32,426	4	10	0,1081	32,426	-22,43	15,5099	4	10	0,1081	32,426	-22,4	15,51
6	5	9	0,1976	59,267	5	9	0,1976	59,267	-50,27	42,6334	5	9	0,1976	59,2667	-50,3	42,63
7	6	40	0,2508	75,226	6	40	0,2508	75,226	-35,23	16,495	6	40	0,2508	75,2257	-35,2	16,5
8	7	51	0,2182	65,473	7	51	0,2182	65,473	-14,47	3,19944	7	51	0,2182	65,4734	-14,5	3,199
9	8	45	0,1247	37,397	8	45	0,1247	37,397	7,6034	1,54589	8	45	0,1247	37,3966	7,603	1,546
					9-10	65	0,0486	14,586	50,414	174,253	9	33	0,0422	12,6578	20,34	32,69
10	9	33	0,0422	12,658		Наблюдаемое значение ста				**518,838**						**1107**
11	10	32	0,0064	***1,928***		Критическое значение стат				11,0705	=ХИ2.ОБР.ПХ(0,05;5)					
n=S(ni)=		300	1,0000	300			Pv=	7E-110	=ХИ2.РАСП.ПХ(K14;5)				Pv=	8E-236		12,59
S(Xi*ni)=		1811														
<x>=		6,04				$\chi^2_{\text{кр}} = \chi^2_{\alpha;\ n-r-1} = \chi^2_{0{,}05;\ 7-1-1} = 11{,}07$					12,59	=ХИ2.ОБР.ПХ(0,05;8-1-1)				
p=		0,6		0												
					Так как		$\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$		, то гипотеза H0 отклоняется.							

Видно, что наблюдаемое значение χ^2 было 519, стало 1107, т.е. почти вдвое больше. Правда, в данном случае это не привело к другому итогу, т.к. слишком чудовищна была разница, которая иллюстрируется следующим рисунком.



Сплошная линия соответствует полигону эмпирических частот, а штриховая — теоретических.

<u>На языке MatCalc</u> программа расчёта выглядит так:

```
alp=0.05;
k=[0;1;2;3;4;5;6;7;8;9;10];
nj=[13;17;15;35;10;9;40;51;45;33;32];
NN=sum(nj);  pnj=nj/NN;  // эмпирические частоты
D=[k,pnj];  m=mean(dlaw(D));
n=#k–1;
p=m/n;  message("Оценка вероятности успеха p="+p);
pj=binlaw(n,p).pe(k);   Npj=NN*pj; // теоретические частоты
n1=nj–Npj;
n2=n1^2/Npj;
Q1=[1:#k,k,nj,pj,Npj,n1];  // начальная таблицы
Q2=select(Q1,Npj>=5);
Q3=select(Q1,Npj<5);
```

```
Q4=sum(cols(Q3)); message("Q4=",Q4);
Q=[Q4;Q2]; message("Q=",Q);
n2Q=Q.c(6)^2/Q.c(5);
Q=[Q,n2Q];
Xi2=sum(n2);  message("Xi2=sum((n–Np)^2/(Np))="+Xi2);
SumQ=sum(cols(Q)); message("SumQ=",SumQ);
Xi2Q=SumQ(#SumQ); message("Xi2Q="+Xi2Q+";  Xi2="+Xi2);
Xcr=x2law(n-2).q(1-alp);
message("Xi_cr= x2law("+(n–2)+").q("+(1-alp)+")="+Xcr);
Pv=x2law(n-2).pg(Xi2Q); message("Pv="+Pv);
if(Xi2Q<Xcr) message("H0 принимается на уровне значимости"
+alp);
else message("H0 отклоняется на уровне значимости"+alp);

H=["j","xi","ni","pi","N*pi","ni-Npi","(ni-Npi)^2/Npi"];
message([H;Q;SumQ]);
savetable([H;round([Q;SumQ],4)],"Пример_7.txt","eng");
line(k,nj,red); line(k,Npj,navy);  //полигон частот
axes(); show(); erase();
line(k,sums(nj),red); line(k,sums(Npj),navy);  //Функции распре-
деления
axes(); show(); erase();
savetable([k,nj,Npj],"Пример7(k,nj,Npj).txt","eng");
```

Пример 8. В автобусном парке ежедневно регистрировалось число автобусов, сошедших с линии в течение рабочего дня. Результаты наблюдений над 200 автобусами представлены в таблице.

x_i	0	1	2	3	4	5	≥ 6
n_i	70	78	34	13	4	1	0

С помощью критерия χ^2 проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что случайная величина X, равная числу автобусов, вышедших из строя в течение рабочего дня, распределена по закону Пуассона.

<u>*Решение в Excel*</u>.

	A	B	C	*D*	E	F	G	H
1				*n* =	200		$\hat{\lambda}$ =	1,03
2								
3	*i*	x_i	n_i	p_i	$n{*}p_i$	$n_i — np_i$	$(n_i - np_i)^2$	$(n_i - np_i)^2/np_i$
4	1	0	70	0,35701	71,40139	-1,40139	1,96390	0,02751
5	2	1	78	0,36772	73,54343	4,45657	19,86098	0,27006
6	3	2	34	0,18937	37,87487	-3,87487	15,01461	0,39643
7	4	3	13	0,06502	13,00370	-0,00370	0,00001	0,00000
8	5	4	4	0,01674	3,34845	0,65155	0,42451	0,12678
9	6	5	1	0,00414	0,82815	0,17185	0,02953	0,03566
10								
□ □	□		200	1	200	1,288E-14		**0,8564**
12								
13						**P-value =**	**0,9307**	

Оценка параметра распределения Пуассона в ячейке H1 $\left(\hat{\lambda} = 1{,}03\right)$ вычисляется по формуле

$\hat{\lambda}$ = СУММПРОИЗВ(B4:B9; C4:C9)/E1

Теоретические вероятности p_i в ячейках D4:D8 вычисляются с использованием функции ПУАССОН.РАСП. В ячейке D4 стоит команда

=ПУАССОН.РАСП (B4;H1;0)

Так как случайная величина, распределенная по закону Пуассона, может принимать любые целые значения, то последний интервал ее значений должен быть расширен до бесконечного. По этой причине последняя вероятность p_6 вычисляется как дополнение до 1 суммы остальных вероятностей. В ячейке D9 команда:

=1 – СУММ(D4:D8)

Сумма значений по строкам столбца H дает наблюдаемое значение статистики $\chi^2_{\text{набл}}$ = 0,8564.

P-значение в ячейке G13 вычисляется по формуле

P-value = ХИ2.РАСП.ПХ(H11; 4)

(Вычисляется вероятность того, что случайная величина, распределенная по закону χ^2 с 4 степенями свободы, принимает значение больше, чем найденное наблюдаемое значение статистики. Число степеней свободы вычисляется как число интервалов (6 – количество различных значений в выборке) минус число параметров, оцениваемых по выборке (1 – параметр $\hat{\lambda}$) и минус 1: 6 – 1 – 1 = 4).

Среди ожидаемых частот np_i есть значения, меньше 5 (в двух последних строках). Объединяем последние три строки (для этого суммируем последние три значения в колонках C и D для значений n_i и p_i; в остальных колонках этой новой строки вычисляем значения по соответствующим формулам). В результате получаем

i	x_i	n_i	p_i	$n{*}p_i$	$n_i - np_i$	$(n_i - np_i)^2$	$(n_i - np_i)^2/np_i$
1	0	70	0,3570	71,4014	–1,4014	1,9639	0,0275
2	1	78	0,3677	73,5434	4,4566	19,8610	0,2701
3	2	34	0,1894	37,8749	–3,8749	15,0146	0,3964
4	3–5	18	0,0859	17,1803	0,8197	0,6719	0,0391
□		200	1	200	1,421E-14		**0,7331**
					P-value =	**0,6931**	

Здесь P-значение вычисляется по команде

ХИ2.РАСП.ПХ (0,7331; 2).

2) Второй параметр функции вычисляется с использованием нового значения числа интервалов (4): 4 – 1 – 1 = 2.

Р-значение равно 0,6931. Нулевая гипотеза на уровне значимости α = 0,05 принимается.

Решение на языке Python.

```
import scipy.stats as sts
import numpy as np # загрузка модулей
n = 200
observed_frequences = np.array ([70, 78, 34, 13, 4, 1])
# наблюдаемые частоты
X = np.arange (0, 6) # значения случайной величины — от 0 до 5.
lambd = np.sum (observed_frequences*X) /n # оценка параметра lambda
# распределения Пуассона,
# совпадающая с выборочным средним
P = np.zeros(6) # заготовка массива для теоретических вероятностей
poisson_rv = st.poisson(lambd) # функция для работы с распределением Пуассона с параметром lambd
for i in range (5): # вычисление теоретических вероятностей pi
              # функция poisson_rv.pmf(i) вычисляет вероятности
              # по формуле Пуассона
   P[i] = poisson_rv.pmf(i)
P[5] = 1-np.sum(P)  # последнюю шестую вероятность вычисляем
                    # как дополнение до 1 суммы предыдущих
expected_frequences = n*P # вычисляем ожидаемые частоты npi
print(expected_frequences)
    # 71.40139211,  73.54343388,  37.87486845,  13.00370483,
      # 3.34845399,   0.82814673

# Среди значений ожидаемых частот есть значения npi < 5;
      # объединяем последние три значения массивов
      # observed_frequences и expected_frequences
expected_frequences_new = expected_frequences[0:4]
expected_frequences_new[3] = np.sum(expected_frequences[3:6])
observed_frequences_new = observed_frequences[0:4]
observed_frequences_new[3] = np.sum(observed_frequences[3:6])
    # вызываем функцию, осуществляющую
    # расчет по критерию χ2
```

```
chisq, p = sts.chisquare(f_obs = observed_frequences_new,
f_exp = expected_frequences_new, ddof=1)
print('chi2 = ', chisq)
print('p-value =', p)
```

Результат:
chi2 = 0.733098191813
p-value = 0.693122106189

Р-значение равно 0,6931. Нулевая гипотеза на уровне значимости α = 0,05 принимается.

Решение на языке R.

Для решения в R понадобится модуль "vcd". Если он отсутствует, нужно его загрузить командой install.packages.

```
install.packages("vcd")
library("vcd")
n <– 200
observed_frequences <– c(70, 78, 34, 13, 4, 1) # наблюдаемые частоты
X <– c(0, 1, 2, 3, 4, 5) # значения случайной величины
# Используемая далее функция goodfit в качестве первого параметра
# принимает матрицу, состоящую из двух столбцов.
# В первом столбце располагаются наблюдаемые частоты,
# а во втором – соответствующие им значения случайной величины.
M <– cbind(observed_frequences, X) # функция cbind формирует
                                   # матрицу из двух векторов-столбцов
gf <– goodfit(M, type = "poisson", method = "MinChisq")
    # Функция goodfit реализует проверку гипотезы.
    # Параметр type задает гипотетический закон распределения
    # (Пуассона), при задании параметра method = «MinChisq» будет производится расчет по критерию χ2
summary(gf) # вывод результатов расчета
```

Результат:

Goodness-of-fit test for poisson distribution

X^2 df P(> X^2)

Pearson 0.8545418 4 0.93099

P-значение равно 0,93099. Нулевая гипотеза на уровне значимости $\alpha = 0{,}05$ принимается.

При использовании функции goodfit() отсутствует возможность корректировки данных с целью добиться выполнения условия $np_i \geq 5$. Причина этого в том, что функция сама производит вычисления теоретических вероятностей, и на вход ей нужно подавать массив отдельных значений случайной величины.

Пример 9. *Критерий инверсий*.

Проверяется гипотеза H_0 числа $x_1, \ldots, x_n$ образуют случайный набор данных против альтернативы H_1 значения $x_1, \ldots, x_n$ не образуют случайный набор данных.

Рассматривается исходная неупорядоченная выборка. Говорят, что пара x_i, x_j в наборе $x_1, \ldots, x_n$ образует ***инверсию***, если $i < j$, но при этом $x_i > x_j$.

В качестве исходной статистики берется следующая «мера беспорядка»: R_n — общее количество инверсий в наборе $x_1, \ldots, x_n$.

Её характеристики: $\mathrm{E}(R_n) = \dfrac{n(n-1)}{4}$; $Var(R_n) = \dfrac{n(n-1)(2n+5)}{72}$.

Известно, что $\dfrac{R_n - \mathrm{E}(R_n)}{\sqrt{Var(R_n)}} \xrightarrow{d} \mathrm{N}(0;1)$ при $n \to \infty$.

Статистика критерия имеет вид: $Z = \dfrac{R_n - \mathrm{E}(R_n)}{\sqrt{Var(R_n)}}$.

Критическим значением критерия является процентная точка нормального распределения $z_{\alpha/2}$ (α — уровень значимости).

Если вычисленное по выборке наблюдаемое значение статистики $Z_{\text{набл}}$ удовлетворяет неравенству $|Z_{\text{набл}}| > z_{\alpha/2}$, то гипотеза H_0 отклоняется на уровне значимости *a*.

Данный метод применим, если объем выборки $n \geq 70$ (в этом случае статистика Z достаточно близка к нормальному закону распределения).

Пример. Смоделируем выборку объема $n = 80$ псевдослучайных чисел с равномерным распределением в интервале (0;1). Проверим, используя критерий инверсий, гипотезу о случайности полученного набора чисел.

Первые 20 строк листа Excel с решением примера представлены на следующей странице.

	B	C	D	E	F	G	H	I	J	K	L
		Смоделировать выборку из U(0,1). Проверить на случайность.									
2	**X*10^6**	**I_X**						$\boldsymbol{n}$ **=**	**80**		
3	847065,5	69	C3=СЧЁТЕСЛИ(B4:B$82;ТЕКСТ(B3;"<0"))					α =	**0,05**	**Критерий инверсий.**	
4	476048,6	36	<476049	D4=ТЕКСТ(B4<0"))				$z_{a/2}$ =	**1,96**	**=НОРМСТОБР(1-J3/2)**	
5	598767,3	44									
6	279325,8	20					Сумма(I_Y)=	R_n =	**1657**	**=СУММ(C3:C81)**	
7	961748,0	73									
8	351477,6	22						**E(**R_n**)=**	1580	**=J2*(J2-1)/4**	
9	470526,2	33									
10	10181,8	1						**D(**R_n**)=**	14483	**=J8*(2*J2+5)/18**	
11	637274,5	42									
12	422938,1	28									
13	815919,9	58						$\mathbf{Z}_{набл}$ =	**0,64**	**=(J6-J8)/КОРЕНЬ(J10)**	
14	201306,6	15									
15	137919,7	11		$\lvert \mathbf{Z}_{набл} \rvert$ =	0,64	$\le$	1,96	= $\mathbf{z}_{a/2}$	=>	элементы выборки случайны	
16	833580,2	57									
17	94603,5	5									
18	131885,5	9									
19	644661,5	37									
20	937872,5	59									

Кажущаяся простота решения достигается применением команд, формируемых с применением функции ТЕКСТ (B4;"<0"). Эта функция формирует текст условия, где сначала стоит знак неравенства «<», а затем — число, содержащееся в указанной ячейке, округлённое до ближайшего целого. Для демонстрации действия этой функции она вставлена в ячейку D4. Именно из-за такого округления в столбце B расположены не функции СЛЧИС(), которые генерируют псевдослучайные числа U(0;1), а команды 10^6*СЛЧИС().

Результат расчёта таков: количество инверсий 1657, что приводит к значению статистики 0,64 и принятию гипотезы о случайности элементов выборки.

Литература

1. *Браилов А. В.* Лекции по математической статистике.— М.: Финакадемия, 2007.— 172 с.

2. *Браилов А. В., Глебов В. И., Криволапов С. Я., Рябов П. Е.* Теория вероятностей и математическая статистика: Учебник-практикум.— М.; Ижевск: НИЦ «Регулярная и хаотическая динамика»; Институт компьютерных исследований, 2016.— 414 с.

3. *Зададаев С. А.* Математика на языке R.— М.: Прометей, 2018.— 324 с.

4. *Маккинни Уэс.* Python и анализ данных.— М.: ДМК Пресс, 2015.— 482 с.

5. *Соловьев В. И.* Анализ данных в экономике.— М.: Кнорус, 2019.— 497 с.

6. www.matcalc.ru

ГЛЕБОВ Владимир Ильич,
КРИВОЛАПОВ Сергей Яковлевич

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Проверка гипотез с использованием Excel, MatCalc, R и Python

Учебное пособие

Публикуется в авторской редакции
Компьютерная верстка *Е. Осипова*
Дизайн обложки *Т. Середа*

Издательство «Прометей»
119002 Москва, ул. Арбат, д. 51, стр. 1
Тел.: +7 (495) 730–70–69
E-mail: info@prometej.su

Подписано в печать 10.12.2018.
Формат 60×84/16. Объем 5,375 п. л.
Тираж 500 экз. Заказ № 815

ISBN 978-5-907100-66-4

9 785907 100664

В. И. Глебов
С. Я. Криволапов

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Проверка гипотез с использованием Excel, MatCalc, R и Python

Учебное пособие

В пособии приведено около 100 задач, посвященных проверке статистических гипотез. Представлены следующие темы: общие свойства (статистика критерия, мощность критерия, P-значение); критерии случайности, независимости, однородности; критерии согласия (простые гипотезы); критерии согласия (сложные гипотезы) и параметрические гипотезы.

Разобраны примеры решения задач проверки гипотез. При решении примеров использовалась среда Excel, языки MatCalc, R и Python.

Пособие предназначено для студентов, обучающихся по направлениям «Экономика», «Менеджмент» и др.

ISBN 978-5-907100-66-4

9 785907 100664

ИЗДАТЕЛЬСТВО ПРОМЕТЕЙ