Практическая работа 4. Генерация распределений. Проверка определений известных распределений

Юрченков Иван Александрович, ассистент кафедры ПМ

2022-10-19

Постановка задачи

1. **Сгенерировать выборку нормального распределения $Y \sim N(\mu, \sigma^2)$ используя определение центральной предельной теоремы**.

Важно. Здесь и далее во всех заданиях слова *случайная реализация случайной величины, распределенной по какому-либо распределению* обозначают вектор с конечным числом значений, сгенерированный из соотвествующего распределения. То есть, если например $Y \sim N(\mu, \sigma^)$ — случайная реализация нормально распределенной величины с параметрами $\Theta = (\mu,\, \sigma^2)$, то Y— это вектор из K значений $Y = (y_1, y_2, \dots, y_K)$, сгенирированный из нормального распределения с заданными конкретными значениями параметров.

- На основе $n \approx 10 \div 20$ равномерно распределенных случайных реализаций случайных величин образовать новую выборку по определению центральной предельной теоремы.

Если $Y_i \sim U(a_i, b_i),\ i = 1, 2, \dots, n$, где Y_i — случайная реализация равномерно распределенной случайной величины с параметрами $a_i \in \mathbb{R},\ b_i \in \mathbb{R}$, то ожидаемая нормально распределенная величина Y будет найдена как:

$$Y = \sum_{i=1}^{n} Y_i,\ i = 1, 2, \dots, n$$

- Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность нормального распределения по несмещенным точечным оценкам $\hat{\mu}, \hat{\sigma}$.
- Провести тест на нормальное распределение с помощью критерия χ^2-Пирсона. Степени свободы рассчитывать как $k = n$.
- Качественно определить влияние числа сгенерированных равномерно распределенных величин на итоговое качество генерации нормального распределения при помощи взятия 3 тестовых генераций при разных n и проведения теста на распределение.

Для генерации выборок рекомендуется пользоваться встроенными в компьютерные статистические пакеты функциями генерации **равномерно распределённых случайных величин**, которые задаются с помощью параметров границ интервала генерации чисел a и b.

2. **Сгенерировать выборку χ^2-распределения $R \sim \chi^2_k$ используя определение распределения χ^2**.

- На основе Z-оценок случайных реализаций нормально распределенных случайных величин $L_i \sim N(\mu_i, \sigma_i^2)$ образовать новую выборку по определению χ^2-распределения:

$$R = \sum_{i=1}^{n} Z[L_i]^2,\ \ Z[L_i] = \frac{L - E[L]}{\sigma[L]},\ L_i \sim N(\mu_i, \sigma_i^2),\ i = 1, 2, \dots, n$$

- Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность χ_k^2 распределения с $k = n$ степенями свободы.
- Провести тест на χ^2 с помощью критерия χ^2-Пирсона.

Для генерации **нормально распределенных реализаций** случайных величин рекомендуется пользоваться встроенными в статистические пакеты функциями для генерации значений выборки из нормального распределения, которые задаются с помощью параметров математического ожидания μ и стандатрного отклонения σ^2.

3. **Сгенерировать выборку распределения Фишера на основе определения**.

- На основе двух случайных реализаций Y_1, Y_2 случайных величин, распределенных по χ^2-распределению со степенями свободы d_1, d_2 соответственно, сгенерировать выборку, распределенную по распределению Фишера $S \sim F(d_1, d_2)$ в соответствии с определением:

$$S = \frac{Y_1/d_1}{Y_2/d_2},\ S \sim F(d_1, d_2).$$

- Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность $F(d_1, d_2)$ распределения.
- Провести тест на распределение Фишера с помощью критерия χ^2-Пирсона.

Для генерации выборки фиксированного размера из распределения χ^2 рекомендуется пользоваться встроенными в статистические пакеты функциями для генерации случайных выборок из распределения χ^2 с df степенями свободы.

4. **Сгенерировать выборку t-распределения на основе определения**.

- На основе $n \approx 2 \div 8$ случайных реализаций $Y_1, Y_2, \dots, Y_n$ случайных величин, распределенных по стандартному нормальному распределению $Y_i \sim N(0, 1),\ i = 1, 2, \dots, n$, сгенерировать выборку $T \sim t(n)$, распределенную по t-распределению Стьюдента с $df = n$ степенями свободы в соответствии с определением:

$$T = \frac{Y_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}}, \quad Y_0 \sim N(0, 1).$$

- Реализовать вычисление аналитической плотности t-распределения Стьюдента с использованием бета-функции:

$$p_t(x \,|\, n) = \frac{1}{\sqrt{n}\,\mathrm{B}(\frac{1}{2}, \frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

где

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt,$$

определённая при Re $x > 0$, Re $y > 0$.

- Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность $t(n)$.
- Для получившейся выбрки провести тест на t-распределение Стьюдента с помощью критерия χ^2-Пирсона, используя в качестве функции вероятности распределения $P_t(x \mid n)$:

$$P_t(x \mid n) = \int_{-\infty}^{x} p_t(z \mid n) dz.$$

5. Для всех заданий количество генерируемых значений выборки установить равным $N \approx 100 \div 1000$. Уровень надежности для критерия χ^2-Пирсона или метода анаморфоз $\gamma = 0.95$.

Пример генерации распределений

1. Генерация нормального распределения из суммы случайных реализаций равномерно распределенной случайной величины

Центральная предельная теорема напрямую утверждает о том, что случайная величина, составленная в виде суммы $S = Y_1 + Y_2 + \ldots Y_n$ сулчайных величин Y_i с конечным математическим ожиданием μ и дисперсией σ^2, обладает свойством:

$$\frac{S - n \cdot \mu}{\sqrt{n} \cdot \sigma} \to N(0,1), \quad n \to +\infty,$$

где $N(0,1)$ — стандартное нормальное распределение.

В другой формулировке теоремы говорится о сумме величин с конечным **неодинаковым** математическим ожиданием μ_i и стандарным отклонением σ_i для каждого члена суммы $S = \sum_{i=1}^{n} Y_i,\ i = 1, 2, \ldots, n.$

ЦПТ Линдеберга

Пусть независимые случайные величины $Y_1, \ldots, Y_n, \ldots$ определены на одном и том же вероятностном пространстве и имеют конечные математические ожидания и дисперсии: $\mathbb{E}[X_i] = \mu_i,\ \mathsf{D}[X_i] = \sigma_i^2.$

Пусть $S_n = \sum\limits_{i=1}^{n} X_i.$

Тогда $\mathbb{E}[S_n] = m_n = \sum\limits_{i=1}^{n} \mu_i,\ \mathsf{D}[S_n] = s_n^2 = \sum\limits_{i=1}^{n} \sigma_i^2.$

И пусть выполняется условие Линдеберга:

$$\forall \varepsilon > 0,\ \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{E}\left[\frac{(Y_i - \mu_i)^2}{s_n^2} \mathbf{1}_{\{|X_i - \mu_i| > \varepsilon s_n\}}\right] = 0,$$

где $\mathbf{1}_{\{|X_i - \mu_i| > \varepsilon s_n\}}$ функция — индикатор.

Тогда

$$\frac{S_n - m_n}{s_n} \to N(0,1)$$

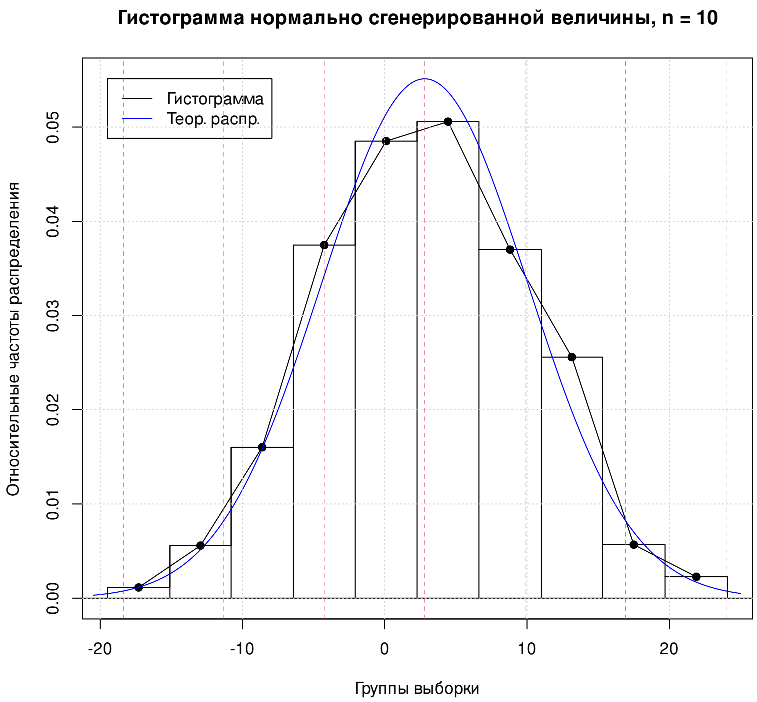
по распределению при $n \to \infty$.

В простыми словами, данная формулировка ЦПТ говорит о том, что если сумма математических ожиданий квадратов z-оценок случайных величин $Y_i,\ i = 1, 2, \ldots n$ в определенной окрестности в предеде $n \to +\infty$ стремится к нулю, то составленная случайная величина:

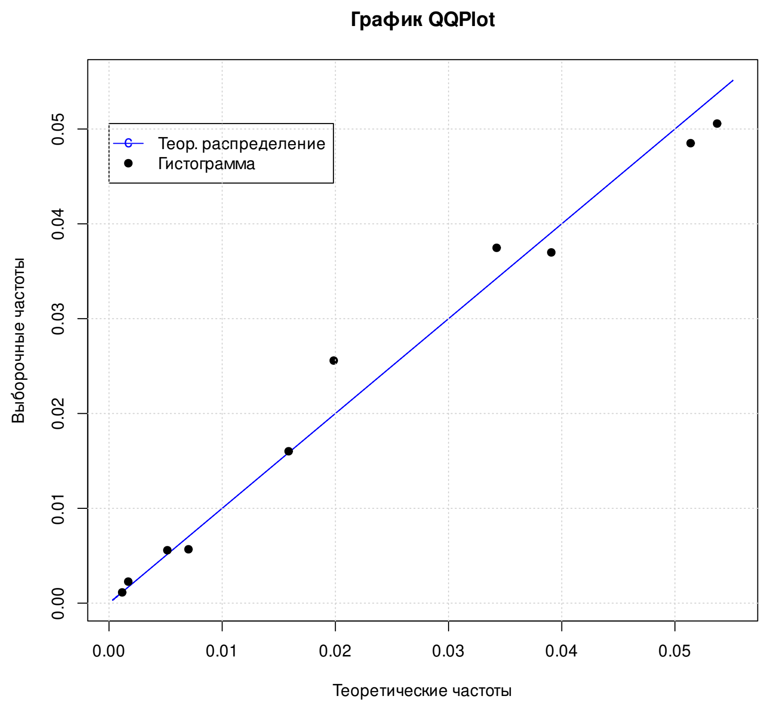
$$S = \sum_{i=1}^{n} Y_i \sim N\left(\mu = \sum_{i=1}^{n} \mu_i, \sigma^2 = \sum_{i=1}^{n} \sigma_i^2\right),$$

будет распределена нормально с математическим ожиданием $\mu = \sum_{i=1}^{n} \mu_i$ и стандартным отклонением $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.

Сгенерировав n равномерных распределений $Y_i \sim Uniform(a_i, b_i),\ (a_i, b_i)\ in\mathbb{R}$ распределений, отобразим гистограмму полученной суммы их реализаций как отдельной случайной величины. На итоговом графике также отобразим μ и $\mu + (-3\sigma, -2\sigma, \dots, 3\sigma)$ значения, полученные напрямую из определения выше, изобразив их штриховой линией.



Полученную гистограмму по её оцененным μ и σ^2 отобразим в спрямляющих координатах нормального распределения, где по оси x отложены теоретические значения вероятности получения тех же значений, что и в исходной гистограмме, а по оси y отобразим сгенерированные значения относительных частот при тех же значениях выборки и биссекрису.



По полученному спрямлению имеем возможность оценить близость полученных зависимостей на основе коэффициента детерминации, посчитав его относительно теоретической зависимости по оцененным параметрам $\hat{y} \sim N(\mu, \sigma^2)$ и практически полученных значений относительных частот, деленных на ширину интервалов по гистограмме $y_i = p_i/h$:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{g} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{g} (\hat{y}_i - E[\hat{y}])^2}.$$

Полученное значение коэффициента $R^2(y, \hat{y}) =$ 0.98, что можно расценивать как положительный тест на нормальное распределение.

2. Генерация χ^2-Распределения по определению

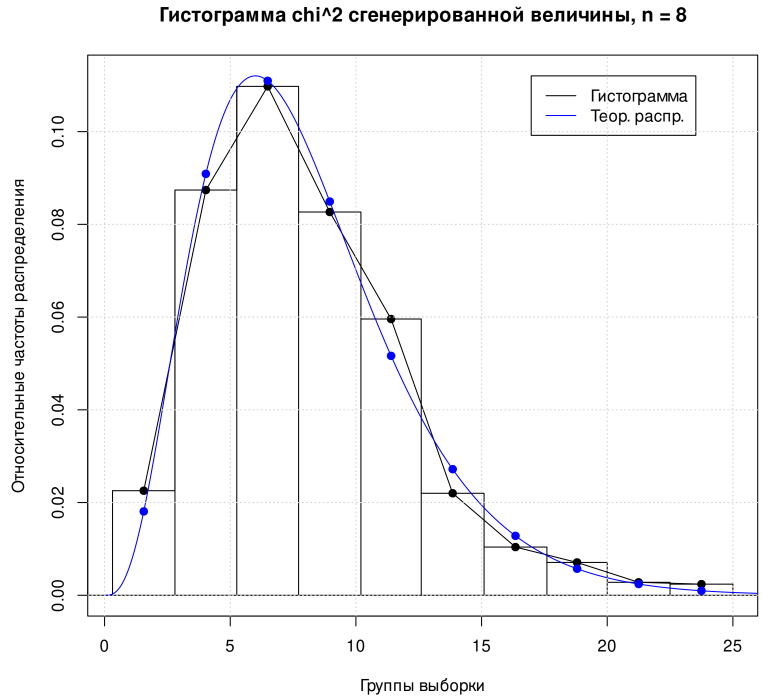
Выборку реализации случайной величны распределенной по χ^2-распределению можно получить из его определения:

$$S = \sum_{i=1}^{n} Z[Y_i]^2, \quad Y_i \sim N(\mu_i, \sigma_i^2), \quad i = 1, 2, \dots, n,$$

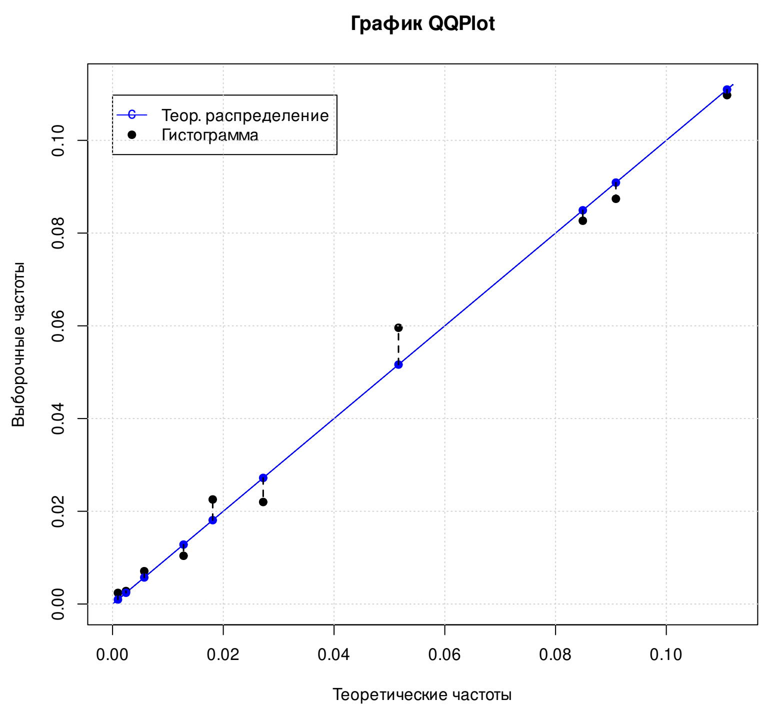
где $Z[Y_i]$ — это Z-оценки соответсвующей реализации нормально распределенной случайной величины $Y_i \sim N(\mu_i, \sigma_i^2)$.

Таким образом выборку S мы можем получить сгенерировав n выборок из нормального распределения $N(\mu_i, \sigma_i^2)$ со случайными параметрами $\Theta = (\mu_i, \sigma_i^2)$, получив из Z-оценки и сложив квадраты полученных значений выборок между собой соответственно.

Проделав такую процедуру получим гистограмму сгенерированной выборки, на которую наложим визуализацию теоретических значений распределения χ^2 со степенями свободы $df = n$.



Изобразим также этот график в новых координатах. Отложим по оси x значения теоретической вероятности из функции плотности χ^2-распределения. По оси y отложим значения полученных относительных частот сгенерированной выборки. Получим спрямление, относительно линейности точек гистограммы которого можно судить о принадлежности выборки распределению.



Коэффициент детерминации между значениями теоретического распределения и значениями частот гистограммы: $R^2(y, \hat{y}) = 0.91$. Значение является довольно высоким, что можно расценивать как положительный тест.

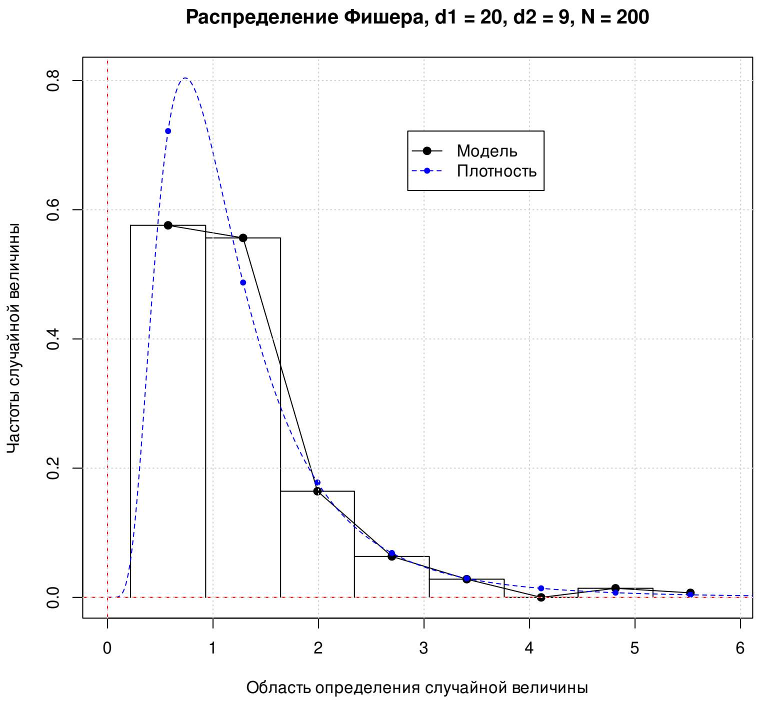
3. Реализация распределения Фишера по определению

Распределение Фишера по определению является отношением реализаций двух случаных величин из χ^2-распределения:

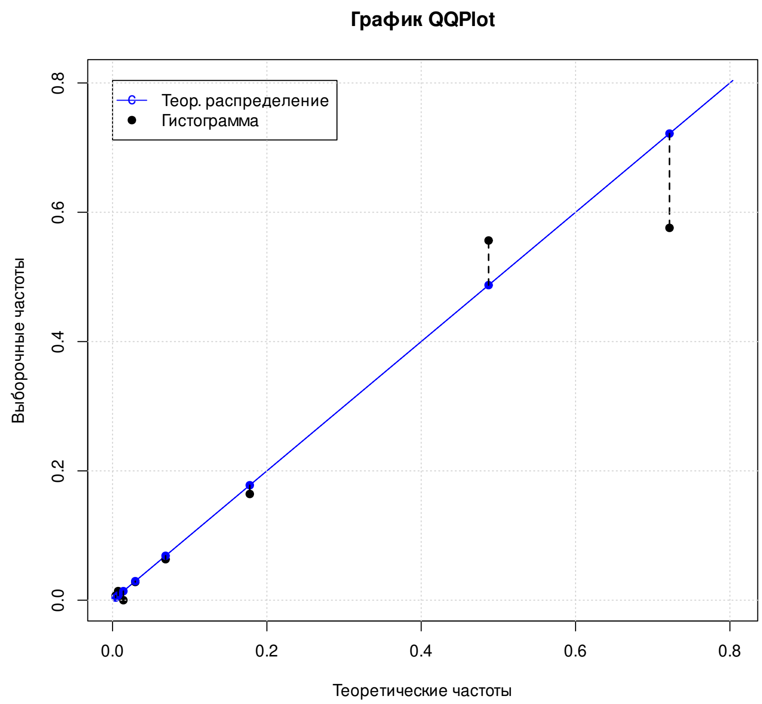
$$S = \frac{Y_1/d_1}{Y_2/d_2} \sim F(d_1, d_2),\ \ Y_1 \sim \chi^2_{d_1},\ Y_2 \sim \chi^2_{d_2}.$$

Таким образом, можно получить выборку, распределенную по распределению Фишера $S \sim F(d_1, d_2)$ с d_1 и d_2 степенями свободы распределений выборок Y_1 и Y_2 соответственно.

Сгенерируем выборку на основе определения распределения Фишера при попощи сгенерированных выборок из распределения χ^2 по N значений с разными степенями свободы $d_1 = 20, d_2 = 9$:



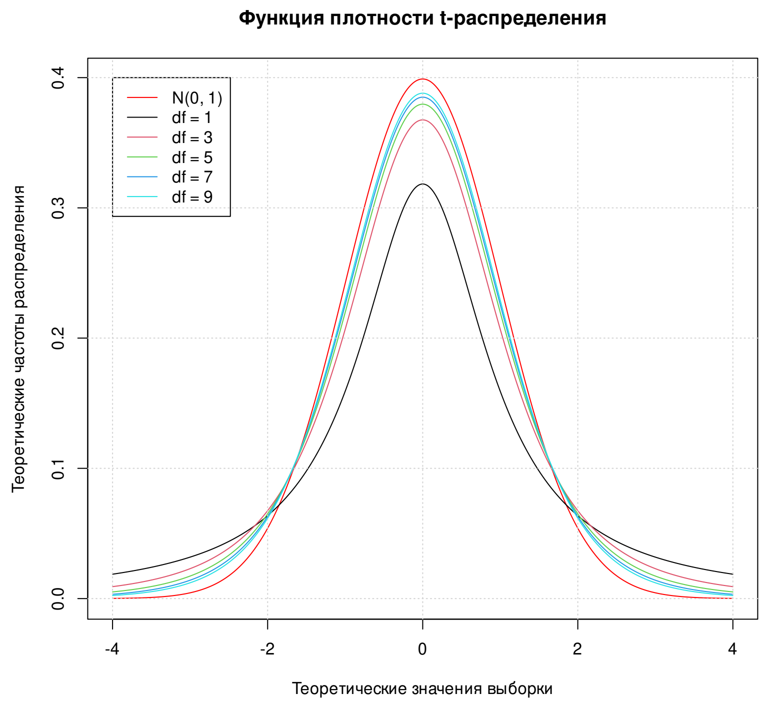
Построим график в спрямляющих координатах по тому же принципу, что и раньше, по оси абсцисс откладываем значения теоретической плотности распределения Фишера, полученной при помощи встроенной в статистический пакет функции, а по оси ординат откладываем сгенерированную выборку распределения Фишера с d_1 и d_2 степенями свободы.



По спрямлению можно оценить коэффициент детерминации между прямой и данными и понять насколько зависимость близка к линейной, что будет говорить о принадлежности выборки к полученному распределению: $R^2(y, \hat{y}) = 0.95$. Значение является довольно высоким, что можно расценивать как положительный тест.

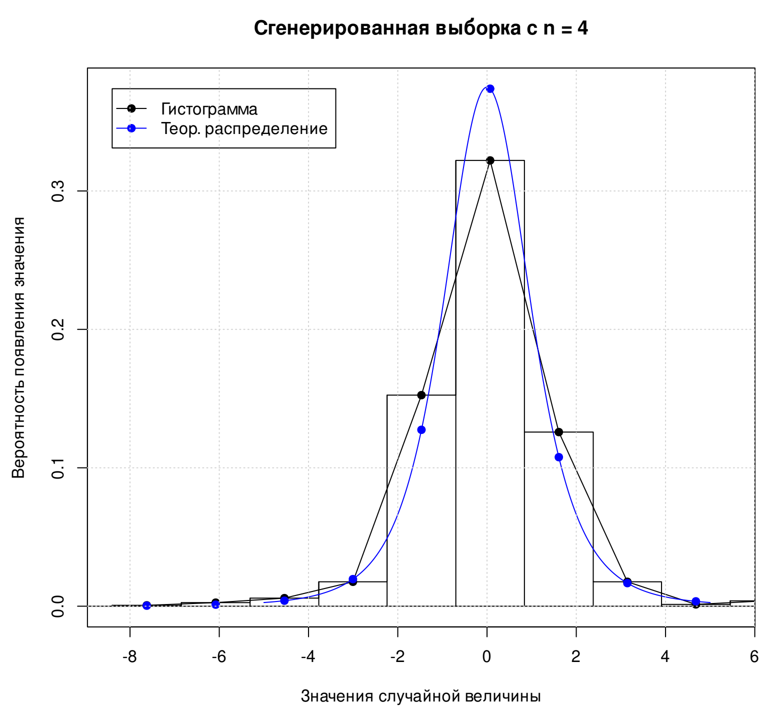
4. Реализация распределения Стьюдента по определению

Реализовав вычисление функции плотности для t-распределения Стьюдента по формулам, мы имеем возможность отобразить на графике полученные нами теоретические плотности распределения при разных степенях свободы.

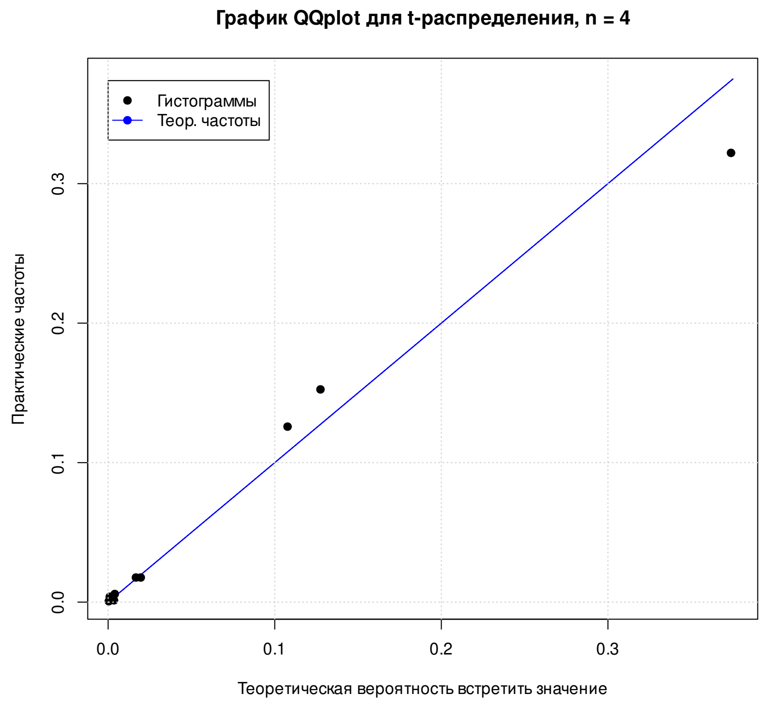


На графике выше мы можем наблюдать сходимость графика плотности t-распределения к стандартному нормальному распределению при увеличении числа степеней свободы. Отсюда можно качественно делать вывод о правдоподности аналитического определения плотности t-распределения по формулам.

Попробуем сгенерировать выборку, удовлетворяющую t-распределению Стьюдента с n степенями свободы. Для этого произведем генерацию нормальных случайных величин в количестве равном n с выборками количеством равным N. Далее для еще одной сгенерированной стандартной нормальной случайной величины произведем вычисления согласно формулам и получим выборку, распределенную по t-распределению с n степенями свободы:



Для полученных теоретических частот и гистограммы для сгенерированной выборки построим график в координатах друг друга $p_i \sim t(n)$ для простой проверки на распределение. Спрямление получим:



И значение коэффициента детерминации: $R^2(y, \hat{y}) = 0.979$. Значение R^2 является довольно высоким, что можно расценивать как положительный тест на распределение Стьюдента.

Вопросы на защиту практической работы

1. Центральная предельная теорема. Реализации случайно распределенных величин. Независимые величины. Степени свободы суммы независимо распределенных величин.
2. Определение нормального распределения. Спрямление для координат нормального распределения. Определение параметров нормального распределения через точечные оценки. Определение параметров нормального распределения, образованного суммой независимых величин, через ЦПТ.
3. Определение распределения Фишера. Аналитические формулы математического ожидания и дисперсии расределения Фишера.
4. t-распределение Стьюдента. Аппроксимации и определение функции плотности. Смесь нормально расределенных величин. Определение Z-оценок.