УДК 658.562.012.7:004.67

И.С.Телина

ИСПОЛЬЗОВАНИЕ R В СТАТИСТИЧЕСКИХ МЕТОДАХ УПРАВЛЕНИЯ КАЧЕСТВОМ

Язык программирования R, благодаря своим особенностям, позволяет реализовывать любые статистические методы, в том числе и в области управления качеством. Помимо пакетов, специально разработанных под задачи управления качеством, R содержит стандартные функции, реализующие основные статистические методы управления качеством. Статья рассматривает один из методов — проверку статистических гипотез, а также способы визуализации данных на примере оценки влияния топологии на параметры полевого транзистора с барьером Шоттки. В статье приводятся рекомендации по выполнению процедуры оценки статистической значимости влияния вносимых изменений в технологический процесс или конструкцию прибора на его характеристики или на процент выхода годных с использованием языка программирования R.

Ключевые слова: статистический метод, язык программирования R, полевой транзистор с барьером Шоттки, статистическая значимость, проверка статистических гипотез

Введение

Статистические методы широко используются для количественной и качественной оценки процессов с целью управления качеством продукции в различных областях, будь то радиоэлектронное производство или разработка ПО. К ним относятся не только простейшие методы контроля процессов, такие как построение гистограммы или контрольных листков, но и методы, требующие основательной подготовки в области математической статистики. В любом случае необходим инструмент, позволяющий выполнять подобные процедуры с минимальными затратами.

Одним из таких инструментов является язык программирования R [1]. Помимо непосредственно языка, R также является свободно распространяемой программной средой для выполнения статистических вычислений и построения графических объектов. К основным достоинствам R можно отнести его открытость, бесплатность, темпы развития и простоту освоения. В R реализованы все методы анализа данных и современные способы визуализации данных [2].

Целью данной статьи является исследование возможностей R для применения статистических методов управления качеством на радиоэлектронном производстве, а именно при проверке статистических гипотез, построении и анализе некоторых статистических графиков.

1. Исходные данные

Исходными данными для анализа послужили результаты измерения двух партий полевых транзисторов с барьером Шоттки (ПТШ). Были получены значения следующих параметров ПТШ на постоянном токе и некоторых параметров, характеризующих шумовые и усилительные свойства:

— начальный ток стока, мА;
— напряжение осечки, В;
— ток утечки затвора, мА;
— крутизна, мА/В;
— коэффициент шума на частоте 50 ГГЦ, дБ;
— коэффициент усиления по мощности на частоте 50 ГГЦ, дБ.

Обе партии транзисторов были изготовлены по одинаковой технологии, но с конструктивными особенностями. Задача исследования — используя язык программирования R, показать, повлияло ли изменение топологии на основные параметры ПТШ. Другими словами необходимо определить, есть ли статистически значимые различия между двумя группами измерений.

Посмотрим на структуру исходных данных.

```
> str(data)
'data.frame':   348 obs. of  7 variables:
 $ Id : num  205 210 200 150 200 200 200 185 175 175 ...
 $ Vth: num  1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 ...
 $ Igl: num  0.041 0.036 0.033 0.063 0.043 0.027 0.03 0.033 0.041 0.054 ...
 $ S  : num  195 200 195 210 200 210 200 190 200 200 ...
 $ NF : num  0.64 0.69 0.69 0.69 0.69 0.7 0.8 0.72 0.72 0.73 ...
 $ G  : num  18.5 19.2 19.1 18.9 18.9 18.8 19.5 19.1 19 19.3 ...
 $ Lot: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Таблица данных содержит результаты измерений 348 полевых транзисторов по шести параметрам:

— Id — начальный ток стока, мА;

— Vth — напряжение осечки, В;

— Igl — ток утечки затвора, мА;

— S — крутизна, мА/В;

— NF — коэффициент шума на частоте 50 ГГц, дБ;

— G — коэффициент усиления по мощности на частоте 50 ГГц, дБ.

Последний седьмой параметр Lot — номер партии. Обе партии изготовлены по стандартной технологии, но вторая — с измененной топологией.

Получим сводную статистику данных.

```
> summary(data)
       Id              Vth              Igl                S               NF               G          Lot
 Min.   :140.0   Min.   :1.000   Min.   :0.02300   Min.   :100.0   Min.   :0.2500   Min.   :18.00   1:174
 1st Qu.:200.0   1st Qu.:1.500   1st Qu.:0.03400   1st Qu.:200.0   1st Qu.:0.6500   1st Qu.:19.00   2:174
 Median :220.0   Median :2.000   Median :0.04200   Median :205.0   Median :0.7300   Median :19.20
 Mean   :226.1   Mean   :1.803   Mean   :0.04372   Mean   :206.4   Mean   :0.7281   Mean   :19.21
 3rd Qu.:240.0   3rd Qu.:2.000   3rd Qu.:0.05200   3rd Qu.:215.0   3rd Qu.:0.8200   3rd Qu.:19.50
 Max.   :400.0   Max.   :2.500   Max.   :0.10000   Max.   :240.0   Max.   :1.1300   Max.   :20.10
```

Для каждого параметра вычислены минимум (Min.), максимум (Max.), медиана (Median), среднее (Mean) и квартили (1st Qu. и 3rd Qu.).

2. Проверка распределений на нормальность

Чтобы визуально оценить распределения измеренных параметров, построим гистограммы. Ниже представлены гистограммы распределения значений начального тока стока и крутизны для двух партий транзисторов, совмещенные с кривыми плотности распределения (рис. 1 и 2). Для построения графиков мы воспользовались возможностями пакета ggplot2 [2].

```
> library(ggplot2)
> ggplot(data, aes(x = Id)) +
    geom_histogram(aes(y = ..density..), binwidth = 30, colour = "black", fill = "white") +
    geom_density(bw = 12, linetype="dashed") +
    facet_grid(.~ Lot) +
    labs(x = "Начальный ток стока, мА", y = "") +
    theme_bw()
> ggplot(data, aes(x = S)) +
    geom_histogram(aes(y = ..density..), binwidth = 10, colour = "black", fill = "white") +
    geom_density(linetype="dashed") +
    facet_grid(.~ Lot) +
    labs(x = "Крутизна, мА/В", y = "") +
    theme_bw()
```

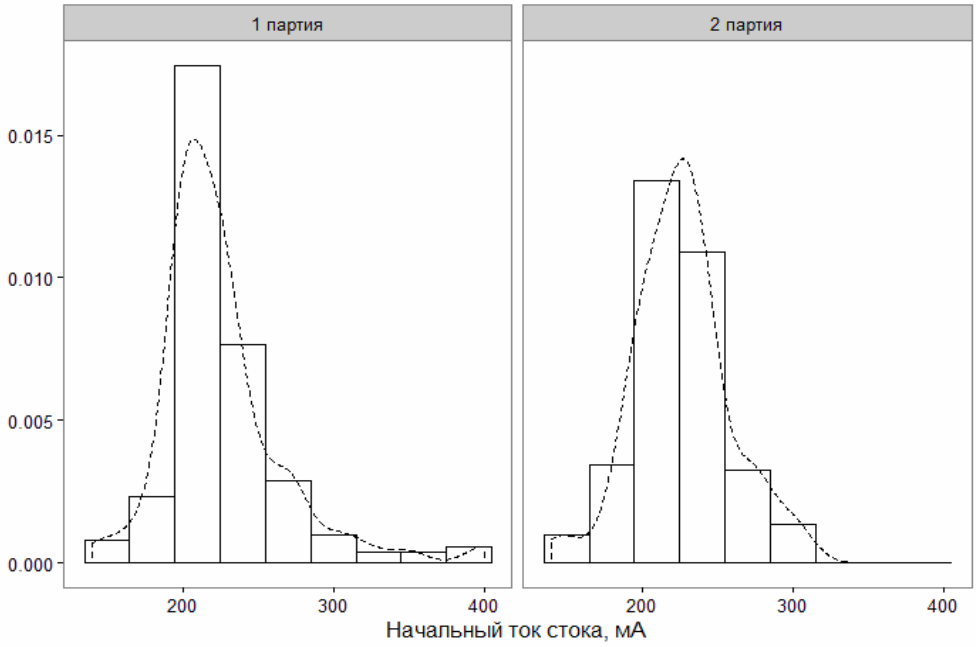

Рис. 1. Распределение значений начального тока стока ПТШ двух партий

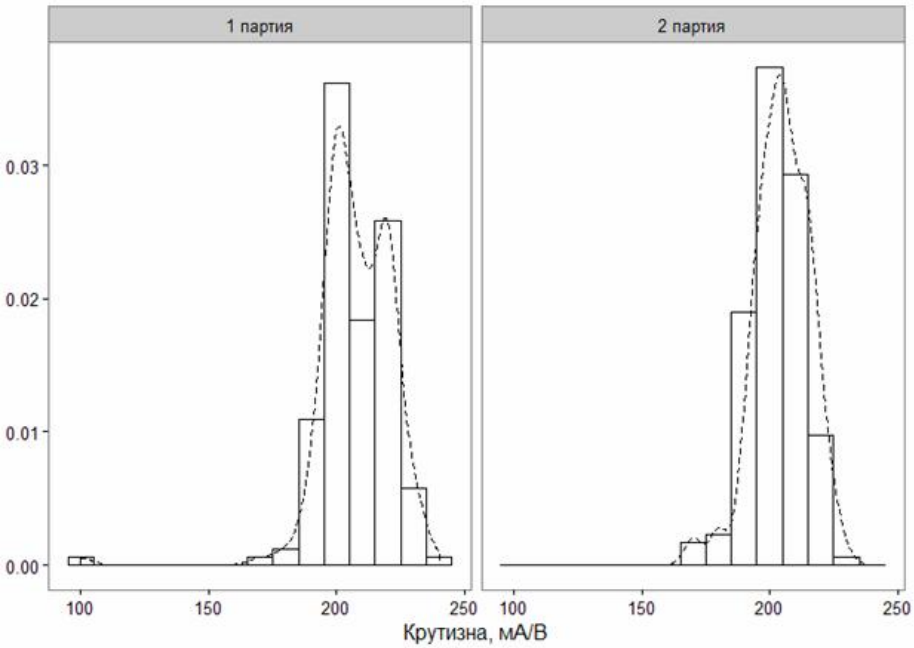

Рис. 2. Распределение значений крутизны ПТШ двух партий

Анализируя полученные изображения, можно отметить, что распределения параметров транзисторов второй партии ближе к нормальному распределению. Проверить данные на нормальность можно несколькими способами. Помимо построения гистограммы, можно наглядно оценить нормальность распределения по графику квантилей (Q-Q plot). Построим его для значений начального тока стока.

```
> qqnorm(data$Id[data$Lot == "1 партия"])
> qqline(data$Id[data$Lot == "1 партия"])
```

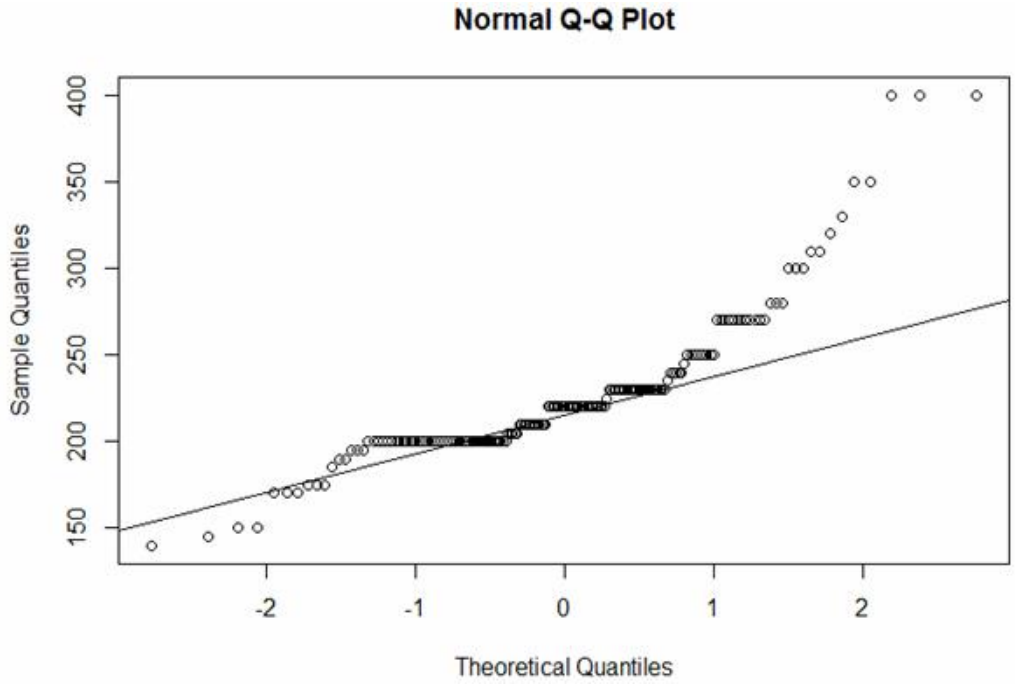

Рис. 3. График квантилей для значений начального тока стока ПТШ первой партии

На графике изображены квантили двух распределений: точками показаны квантили распределения значений начального тока стока, прямая представляет стандартное нормальное распределение. Видно, что анализируемое распределение по сравнению с нормальным имеет несколько больше значений в области малых токов (точки ниже прямой в левой части графика), а также избыток больших значений токов (точки выше прямой в правой части графика). Это подтверждает гистограмма на рис. 1. Приведем также график квантилей распределения значений начального тока стока ПТШ с измененной топологией.

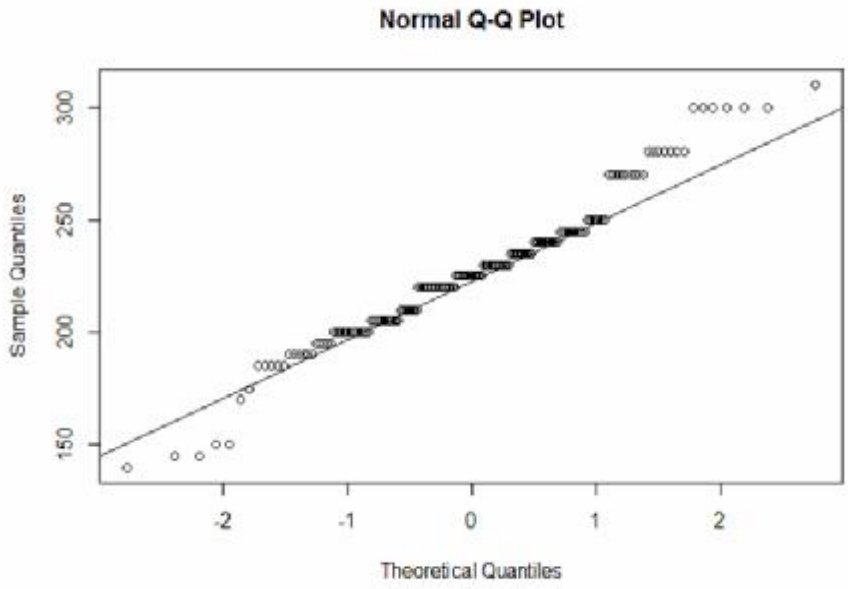

Рис. 4. График квантилей для значений начального тока стока ПТШ второй партии

Рисунок демонстрирует подобный характер эмпирического распределения и его отличия от теоретического нормального распределения.

3. Выявление выбросов

Для нахождения выбросов можно воспользоваться графическим методом, а именно построением диаграммы размаха («ящика с усами»).

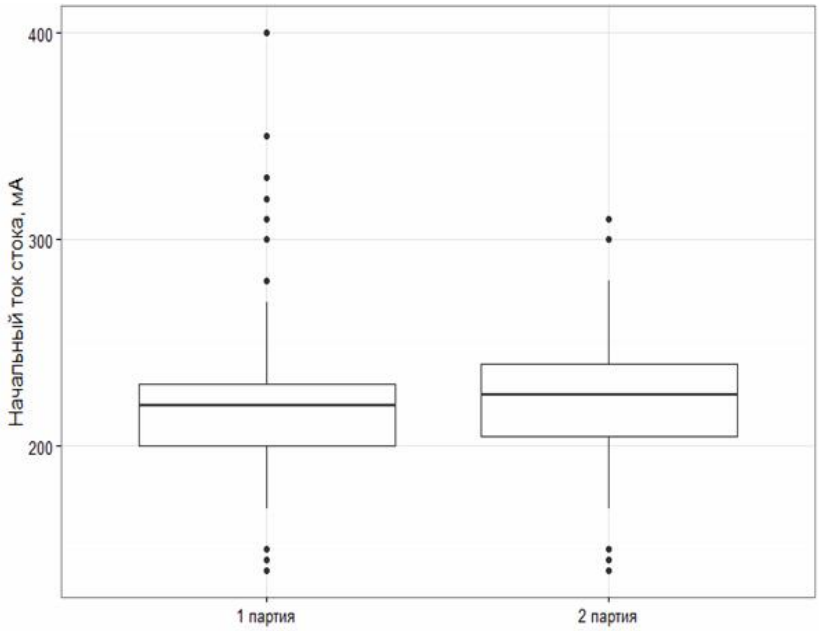

Рис. 5. Диаграмма размаха значений начального тока стока

На диаграмме размаха (рис. 5) показаны медианы двух групп данных (линия внутри прямоугольника) и межквартильный размах — разность между третьим и первым квартилями (высота прямоугольника). Точки, лежащие выше и ниже «усов ящика» не попадают в диапазон значений:

[(Q1 — 1.5 * (Q3 — Q1)); (Q3 + 1.5 * (Q3 — Q1))], (1)

где Q1 — первый квартиль, Q3 — третий квартиль.

Такие точки являются потенциальными выбросами. Эти значения лучше удалить из выборки перед применением некоторых тестов.

В отдельных случаях точки, лежащие вне диапазона (1), будут являться не выбросами, а, частью распределения, отличного от нормального. К примеру, рассмотрим более подробно распределение значений

токов утечки затвора ПТШ. Построим для этих данных и диаграмму размаха (рис. 6), и гистограмму (рис. 7).

```
> ggplot(data[data$Lot == "1 партия", ], aes(x = factor(0), y = Igl)) +
    geom_boxplot() + xlab("") + ylab("Ток утечки затвора, мА") +
    scale_x_discrete(breaks = NULL) + coord_flip() +
    theme_bw()
> ggplot(data[data$Lot == "1 партия", ], aes(x = Igl)) +
    geom_histogram(aes(y=..density..), binwidth=0.005, colour="black", fill="white") +
    geom_density(linetype="dashed") +
    labs(x = "Ток утечки затвора, мА", y = "") +
    theme_bw()
```

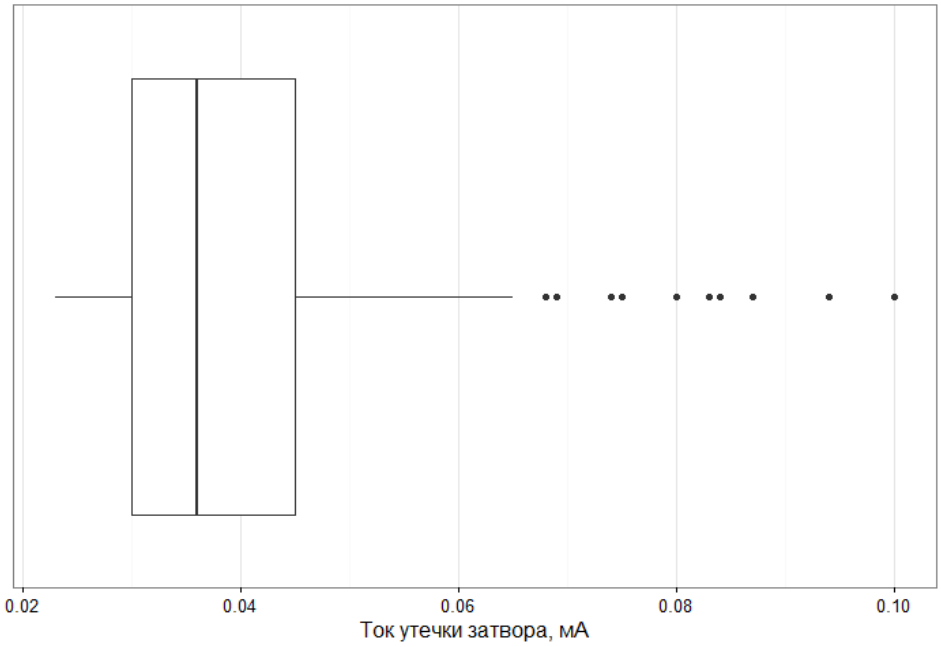

Рис. 6. Диаграмма размаха значений тока утечки затвора для первой партии ПТШ

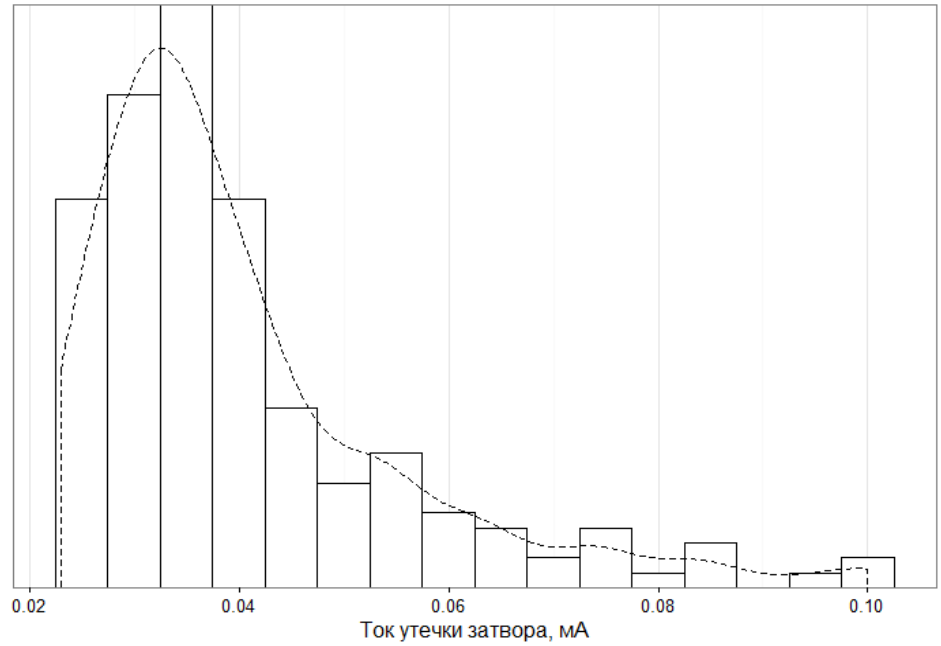

Рис. 7. Гистограмма значений тока утечки затвора для первой партии ПТШ

Как видно из рис. 7, данное распределение не является нормальным. Это доказывают результаты теста Шапиро-Уилка [3].

```
> shapiro.test(data$Igl[data$Lot == "1 партия"])
        Shapiro-Wilk normality test
data:  data$Igl[data$Lot == "1 партия"]
W = 0.83258, p-value = 7.439e-13
```

Полученная вероятность ошибки p равна 7.439e-13, что намного меньше стандартного уровня значимости теста, принятого равным 0.05. Это позволяет отклонить нулевую гипотезу о нормальности изучаемого распределения. Распределение значений токов утечки больше напоминает логнормальное или бета-распределение. Поэтому значения, показанные точками справа от «ящика с усами» на рис. 6, идентифицировать как выбросы не будем.

4. Сравнение двух групп данных

Диаграммы размаха также используются для визуального сравнения сгруппированных данных. Построим диаграмму размаха для значений коэффициента шума двух партий ПТШ (рис. 8).

```
> ggplot(data, aes(x = Lot, y = NF)) +
    geom_boxplot() +
    labs(x = "", y = "Коэффициент шума на частоте 50 ГГц, дБ") +
    theme_bw()
```

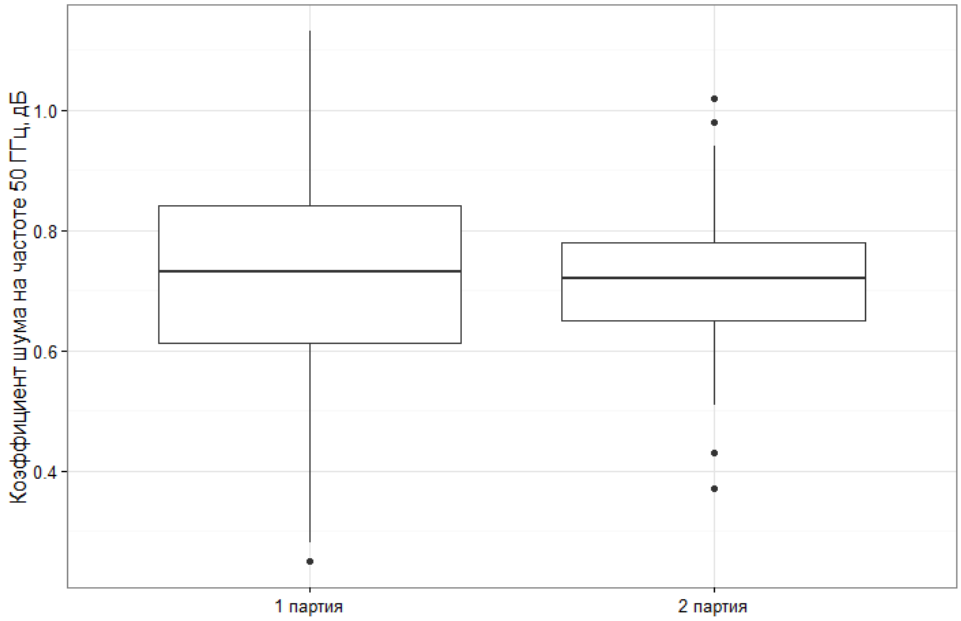

Рис. 8. Диаграммы размаха значений коэффициента шума двух партий ПТШ

Визуально между медианами двух групп практически нет различий. Можно отметить, что значения коэффициента шума транзисторов второй партии имеют меньший размах, что подтверждают построенные ниже гистограммы распределений (рис. 9).

```
> ggplot(data, aes(x = NF)) +
    geom_histogram(aes(y=..density..), binwidth=0.1, colour="black", fill="white") +
    geom_density(linetype="dashed") +
    labs(x = "Коэффициент шума на частоте 50 ГГц, дБ", y = "") +
    theme_bw() + facet_grid(.~ Lot)
```

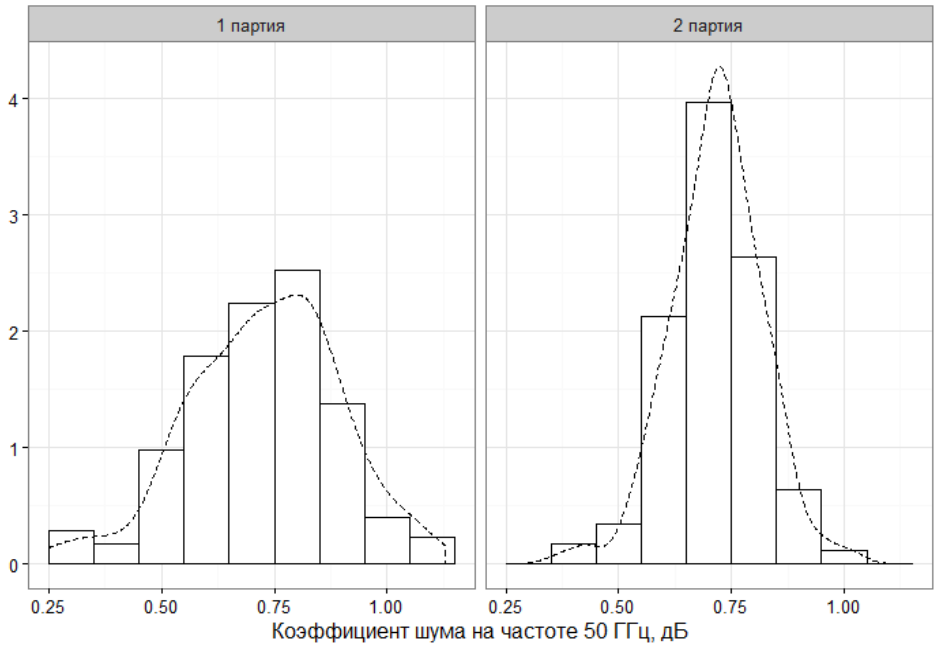

Рис. 9. Гистограммы распределений значений коэффициента шума двух партий ПТШ

Проверим данные распределения на нормальность.

```
> by(data$NF, INDICES = data$Lot, shapiro.test)
data$Lot: 1 партия
        Shapiro-Wilk normality test
data:  dd[x, ]
W = 0.98928, p-value = 0.2131
------------------------------------------------------------------------------
data$Lot: 2 партия
        Shapiro-Wilk normality test
data:  dd[x, ]
W = 0.99, p-value = 0.262
```

Согласно тесту Шапиро-Уилка, оба p-значения больше 0.05, что не позволяет нам отклонить нулевую гипотезу о нормальности распределений.

Проверим условие однородности дисперсий для двух групп значений коэффициента шума, применив тест Бартлетта [3].

```
> bartlett.test(NF ~ Lot, data)
        Bartlett test of homogeneity of variances
data:  NF by Lot
Bartlett's K-squared = 40.476, df = 1, p-value = 1.991e-10
```

Результаты теста показывают, что внутри групп дисперсия неоднородна (p-значение много меньше 0.05). Это следует учесть при применении теста Стьюдента для проверки равенства средних двух выборок.

```
> t.test(data$NF[data$Lot=="1 партия"], data$NF[data$Lot=="2 партия"], var.equal=FALSE)
        Welch Two Sample t-test
data:  data$NF[data$Lot == "1 партия"] and data$NF[data$Lot == "2 партия"]
t = 0.12045, df = 286.15, p-value = 0.9042
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02733109  0.03089431
sample estimates:
mean of x mean of y
0.7290230 0.7272414
```

Согласно результатам теста Стьюдента мы не можем отклонить нулевую гипотезу о равенстве средних значений двух генеральных совокупностей (p-значение больше 0.05). Другими словами, статистически значимых различий между средними значениями коэффициента шума двух партий ПТШ не обнаружено.

Рассмотрим теперь распределение последнего параметра ПТШ — коэффициента усиления. На рис. 10 представлены гистограммы распределений, на рис. 11 — диаграммы разброса значений коэффициентов усиления ПТШ двух партий.

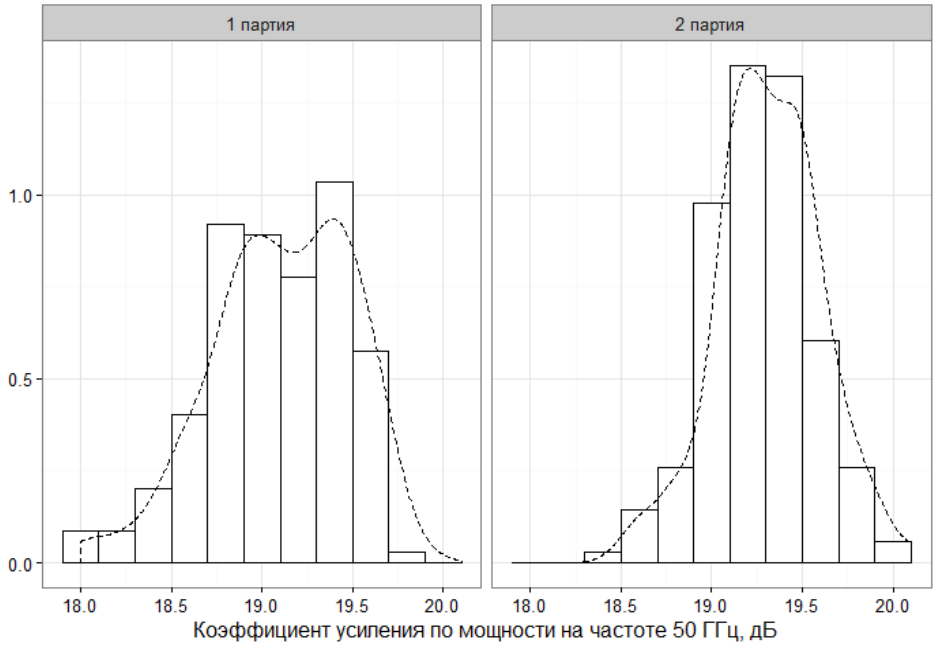

Рис. 10. Гистограммы распределений значений коэффициента усиления двух партий ПТШ

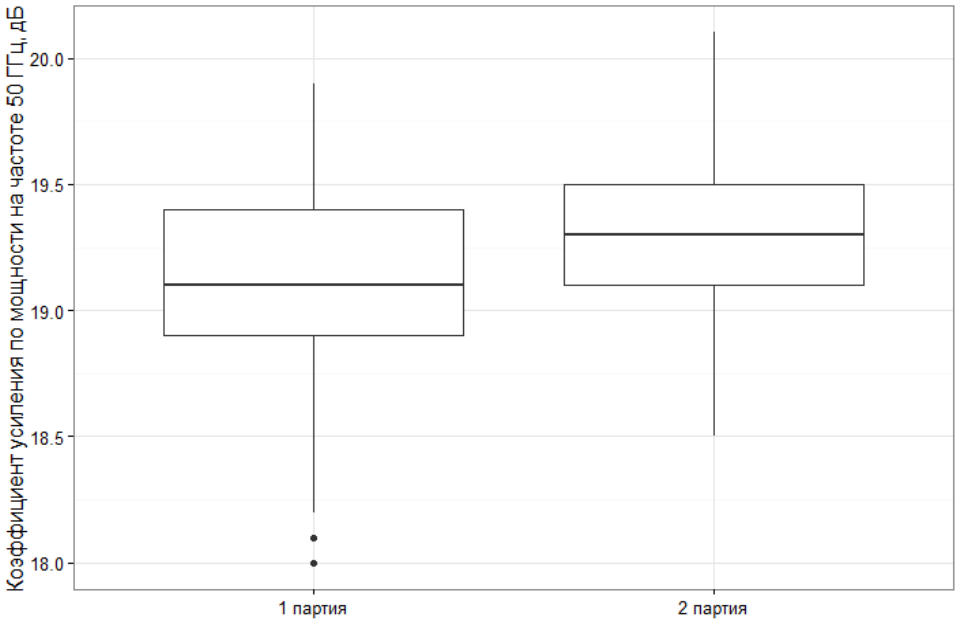

Рис. 11. Диаграммы разброса значений коэффициента усиления двух партий ПТШ

Наблюдается небольшая разница между группами значений коэффициента усиления. Для того, чтобы воспользоваться формальными тестами для сравнения двух групп данных, необходимо убедиться, что данные распределены нормально. Выполним тест Шапиро-Уилка для каждой группы данных.

```
> by(data$G, INDICES = data$Lot, shapiro.test)
data$Lot: 1 партия
        Shapiro-Wilk normality test
data:  dd[x, ]
W = 0.9691, p-value = 0.0006518
```

```
------------------------------------------------------------------------------
data$Lot: 2 партия
        Shapiro-Wilk normality test
data:  dd[x, ]
W = 0.98263, p-value = 0.02879
```

P-значения обоих тестов меньше 0.05. Требование нормальности распределения данных не выполняется, поэтому воспользуемся непараметрическим тестом Уилкоксона для сравнения двух групп данных (он же U-критерий Манна-Уитни) [3].

```
> wilcox.test(data$G[data$Lot == "1 партия"], data$G[data$Lot == "2 партия"])

        Wilcoxon rank sum test with continuity correction

data:  data$G[data$Lot == "1 партия"] and data$G[data$Lot == "2 партия"]

W = 9958, p-value = 3.133e-08

alternative hypothesis: true location shift is not equal to 0
```

Данные двух групп статистически значимо отличаются друг от друга (p-значение много меньше 0.05). Это означает, что изменение в топологии при изготовлении ПТШ второй партии привело к изменению коэффициента усиления по мощности транзистора.

Заключение

В производственной практике довольно часто встречаются задачи оценки влияния вносимых изменений в технологический процесс или конструкцию прибора на его характеристики или процент выхода годных. Лишь использование статистических методов позволяет понять, произошло ли ожидаемое изменение в параметрах прибора и повлияло ли оно на процент выхода годных, или же мы имеем дело со случайным отклонением. Надо отметить, что на практике зачастую такая проверка просто опускается.

В статье рассмотрены возможности языка программирования R при проверке статистических гипотез и визуализации полученных данных на примере оценки влияния топологии на параметры полевого транзистора с барьером Шоттки. Проверив статистическую гипотезу о равенстве средних значений генеральных совокупностей по двум выборкам, мы показали, что изменение топологии оказало статистически значимое влияние лишь на значения коэффициента усиления по мощности. Были проверены статистические гипотезы о соответствии эмпирических распределений нормальному для нескольких параметров ПТШ, а также проверена статистическая гипотеза об однородности дисперсий двух групп. Помимо этого, в статье были построены с использованием R и проанализированы гистограммы эмпирических распределений и диаграммы разброса значений некоторых параметров ПТШ.

Статья носит больше методологический характер, чем аналитический, поэтому в заключении приведем общий порядок выполнения оценки статистической значимости влияния вносимых изменений в технологический процесс или конструкцию прибора на его характеристики или на процент выхода годных с использованием R. Весь этот процесс построен практически на одном статистическом методе — проверке статистических гипотез. В табл. приведены некоторые критерии, используемые при проверке соответствующих статистических гипотез, а также предложены функции R. В скобках после наименования функции указан пакет, где можно ее найти.

Таблица

Этапы сравнения двух групп данных

Этап	Критерии	Функции R
1. Загрузка данных в R		read.table, read.csv
2. Вычисление описательных статистик в целом и по группам данных. Визуализация результатов (построение гистограмм, диаграмм размахов)		str, summary, aggregate table, prop.table describe (Hmisc, psych) plot, hist, barplot, boxplot ggplot (ggplot2)
3. Проверка распределений на нормальность. Визуализация результатов	критерий Шапиро-Уилка	shapiro.test qqnorm, qqplot

Этап	Критерии	Функции R
4. Выявление выбросов	— критерий на основе межквартильного размаха — критерий Граббса — критерий Диксона	 grubbs.test (outliers) dixon.test (outliers)
5. Проверка гипотез об однородности дисперсий (сравнение дисперсий генеральных совокупностей по выборкам)	— критерий Бартлетта — критерий Фишера, если данные прошли проверку на нормальность; — критерий Кохрана в остальных случаях	bartlett.test var.test cochran.test (outliers)
6. Сравнение двух групп данных:		
— сравнение количественных данных (сравнение средних величин генеральных совокупностей)	— критерий Стьюдента, если данные прошли проверку на нормальность и/или нет выбросов; — критерий Уилкоксона в остальных случаях или когда данных мало	t.test wilcox.test
— сравнение категориальных данных (анализ таблиц сопряженности)	— критерий согласия Пирсона (хи-квадрат); — точный критерий Фишера, если данных мало	chisq.test fisher.test

1. R-project [Электронный ресурс] URL: https://www.r-project.org/ (дата обращения: 01.08.2016).
2. Кабаков Р. И. R в действии. Анализ и визуализация данных в программе R / Пер. с англ. П.Волковой. М.: ДМК Пресс, 2014. 588 с.
3. Кобзарь А. И. Прикладная математическая статистика. М.: Физматлит, 2006. 816 с.

References

1. R-project. Available at: https://www.r-project.org/ (accessed 01.08.2016).
2. Kabacoff R.I. R in Action. Data analysis and graphics with R. Shelter Island, Manning Publications Co, 2011. 447 p. (Russ. ed.: Kabakov R.I. R v dejstvii. Analiz i vizualizaciya dannyh v programme R. Moscow, DMK Press, 2014. 588 p.).
3. Kobzar' A.I. Prikladnaya matematicheskaya statistika. Moscow, Fizmatlit, 2006. 816 p.

Telina I.S. R usage for statistical quality control methods. R is a programming language for statistical analysis, including methods of quality control. In addition to the packages, designed specially for quality control tasks, R contains standard functions that perform basic statistical methods of quality control. The article deals with one of these methods — statistical hypothesis testing, as well as methods of data visualization using an example of impact assessment of topology on the Schottky-field-effect transistor`s parameters. The article gives some recommendations on the implementation of the impact significance assessment of the changes in technological process or device structure on its characteristics or the yield value using R.

Keywords: statistical method, programming language R, Schottky-field-effect transistor, statistical significance, statistical hypothesis testing.

Сведения об авторе. И.С.Телина — кандидат технических наук, кафедра физики твердого тела и микроэлектроники ИЭИС, irina.telina@novsu.ru.

Статья публикуется впервые. Поступила в редакцию 16.08.2016.