1 Описательная статистика и визуализация

1.1 Выборка и реализация

В рамках курса статистики мы будем работать с набором из n наблюдений $x_1, \ldots, x_n$ некоторого набора случайных величин (векторов) $X_1, \ldots, X_n$. Иначе говоря, мы будем рассматривать н.о.р. случайные величины X_i и при некотором ω будем наблюдать набор значений $(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$, который и будем обозначать $x_1, \ldots, x_n$.

Определение 1. Набор величин $(X_1, \ldots, X_n)$ называется выборкой, а их значений $x_1, \ldots, x_n$ — реализацией.

Пример 1. При подбрасывании симметричной монеты 5 раз выпал набор "орел, решка, решка, орел, орел". Указанный набор является реализацией (что выпало), а пять бернуллиевских величин с параметром 0.5 являются выборкой (что могло).

Мы будем получать из реализации некоторую информацию про распределения величин.

Пример 2. Приведем примеры статистических задач:

- На основе 100 подбрасываний игрального кубика сказать насколько он похож на правильный.
  С точки зрения выборки мы имеем 100 независимых одинаково распределенных величин с 6 возможными значениями с неизвестными вероятностями. Мы хотим понять, насколько вероятно могут ли все грани иметь вероятность $1/6$ (эта постановка несколько бессмысленна, а точную мы дадим позднее).

- На основе сотни геологических замеров найти область, скорее всего содержащую месторождение нефти.
  Чтобы перейти к математически содержательной задаче, мы должны сформулировать предположения о том, как замеры связаны с наличием вблизи месторождения.

- На основе сравнения семидесяти больных, принимавших лекарство, и ста не принимавших, понять правда ли первая категория более здоровая.
  Наша выборка состоит из 70 наблюдений первого типа и 100 второго, мы хотим понять правда ли вторые координаты в среднем (в каком-то смысле) больше.

- На основе ста наблюдений за числовыми показателями больных до и после приема лекарства, сказать работает ли лекарство.
  Наша выборка состоит из 100 независимых двумерных векторов с зависимыми координатами, мы хотим понять правда ли разность первой и второй координат векторов в среднем (в каком-то смысле) положительна.

- На основе многократного измерения детали оценить ее объем.
  Чтобы перейти к математической задаче нужно представлять как устроена погрешность измерения.

- На основе пятидесяти лет наблюдения за погодой предсказать погоду на следующей неделе.
  Чтобы перейти к математически содержательной задач, нужно представлять как связаны друг с другом разные дни.

1.2 Описательная статистика

Начнем с так называемой описательной статистики (descriptive statistics)

Реализацию можно охарактеризовать несколькими числовыми характеристиками. Наиболее очевидные два параметра: $\min X_i$ и $\max X_i$ — границы в которых изменяются наблюдения. Этот промежуток можно дополнить другими характеристики (позже мы назовем их "выборочными квантилями"):

Определение 2. *Медианой* реализации $x_1, \ldots, x_n$ называют величину

$$MED = \begin{cases} x_{(k)}, & n = 2k, \\ (x_{(k)} + x_{(k+1)})/2, & n = 2k+1. \end{cases}$$

Здесь $x_{(1)}, \ldots, x_{(n)}$ — вариационный ряд, то есть упорядоченные по возрастанию наблюдения.

Определение 3. *Нижним выборочным квартилем* и *верхним выборочным квартилем* называют, соответственно,

$$q_{0.25} = x_{([(n+1)/4])}, \quad q_{0.75} = x_{([3(n+1)/4])}.$$

Определение 4. *Интерквантильным размахом* называют разницу квартилей

$$q_{0.75} - q_{0.25}$$

Медиана разделяет выборку на две половины, в каждой из которых одинаковое количество наблюдений. Нижний и верхний квартиль делят каждую из полученных групп еще на две равные (с точностью до 1 наблюдения) части. Таким образом, глядя на $\min, q_{0.25}, MED, q_{0.75}, \max$ мы видим как расположены 4 группы наблюдений — четверть самых маленьких, четверть побольше, четверть еще побольше и четверть самых больших. Интерквантильный размах характеризует длину интервала, содержащего "средние" две четверти, то есть 50% наблюдений.

Пример 3. Рассмотрим ряд 0, 3, 5, -1, 14, 6, -2. Тогда вариационный ряд для него это -2, -1, 0, 3, 5, 6, 14. Медиана нашего ряда равна 3, нижний квартиль –1, верхний квартиль 6.

Пример 4. Отметим, что медиана болезненно себя чувствует, если примерно половина выборки принимает одни значения, а вторая половина — резко отличающиеся другие. Скажем, если половина работников фирмы получает 100\$, а вторая половина 2500\$, то медиана равна 1300\$. Если мы наймем еще одного работника на малооплачиваемую должность, то медиана тут же просядет до 100\$, а если одного на высокооплачиваемую, то поднимется до 2500\$.

Одно единственное наблюдение, причем совершенно рядовое, существенно изменило медиану. И это случилось бы даже если бы наблюдений было очень много.

Середину выборки можно охарактеризовать не только медианой, но и средним:

Определение 5. *Выборочным средним* $x_1, \ldots, x_n$ называют

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Вопрос 1. Чему равны выборочные среднее и медиана ряда из

Пример 5. Оно из отличий медианы и среднего хорошо демнонстрирует простой пример с зарплатами. Пусть начальник компании получает 1010 000 рублей в месяц, а каждый из 99 работников - 10000 рублей. Медиана дохода при этом 10000 рублей, а среднее — 20 000 рублей. Иногда это приводят как показатель того, что медиана более правильный показатель среднего дохода компании, хотя скорее они просто разные —- медиана характеризует заработок среднего работника, а среднее — сколько получит работник, если все разделить поровну.

Выборочное среднее приближает математическое ожидание нашего ряда (например, при больших n это так в силу ЗБЧ), а для приближения дисперсии используют величину

$$S_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Почему эта величина оценивает дисперсию и как вообще оценивать какие-либо параметры, мы поговорим позже. Пока же отметим, что параметры $\overline{x}$ и S_0 характеризуют среднее и разброс.

1.3 Столбцовая диаграмма, гистограмма и ядерная оценка плотности

Пусть x_i скалярные величины, X_i независимые одинаково распределенные и мы хотим охарактеризовать распределение соответствующей выборки. В дискретном случае мы можем описать распределение с помощью функции масс, а в абсолютно-непрерывном — с помощью функции плотности, однако, обе нам неизвестны.

Давайте попробуем "оценить" их с помощью наших наблюдений.

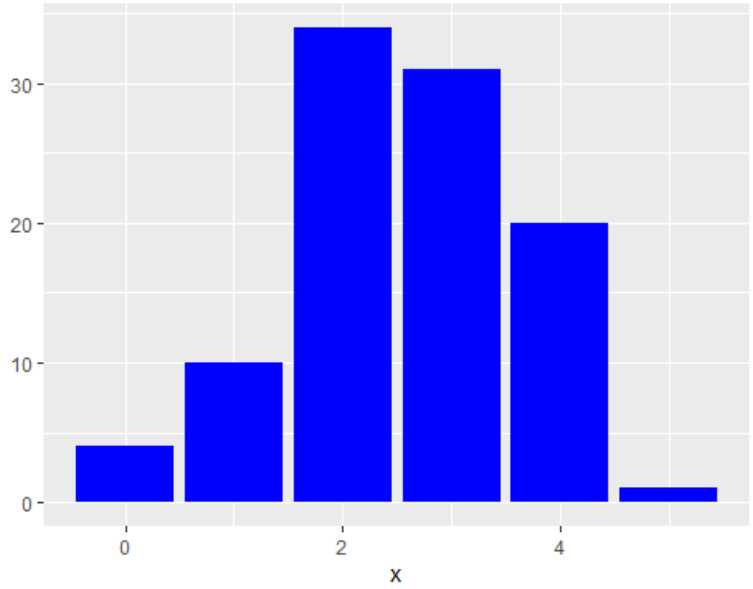
- Для дискретной величины естественно посчитать набор частот $\nu_1, \nu_2, \ldots, \nu_k$ появлений каждого из значений$y_1, \ldots$ нашей величины в реализации. Тогда график, на котором в точке y_i находится столбик высотой ν_i/n, будет

”близок” к графику функции масс, если n велико, поскольку по УЗБЧ п.н.

$$\frac{\nu_i}{n} \to \mathbf{P}(X = y_i),\ n \to \infty.$$

Такой график называется bar plot или столбцовой диаграммой.

Рис. 1: Столбцовая диаграмма для биномиального распределения



- Для абсолютно-непрерывной величины вспомним, что плотность можно задавать соотношением

$$f_X(x) = \lim_{h \to 0} \frac{\mathbf{P}(X \in U_h(x))}{h},$$

где U_h при каждом h есть некоторый интервал, содержащий точку x. Значит, можно разбить область значений величины на отрезки $I_1, \ldots, I_k$ ширины h, найти частоты ν_i попаданий в каждый отрезок:

$$\nu_i = \frac{1}{n} \#\{j : x_j \in I_i\},$$

где $\#$ — число элементов в множестве. Тогда плотность можно оценить функцией, равной ν_i/h при $x \in I_i$. График такой функции называется гистограммой. Это кусочно-постоянная оценка плотности.

Если приблизить функцию масс столбцовой диаграммой можно достаточно точно при большом числе наблюдений, то для оценки плотности нам нужно с увеличением n еще и уменьшать h. При этом при заданном (даже очень большом) n достаточно трудно выбрать подходящее h. Слишком маленькое h приведет к тому, что функция будет равна $1/(nh)$ в окрестности x_i и 0 иначе. Слишком большое — что распределение будет равномерным на отрезке от минимального x_i до максимального x_i.

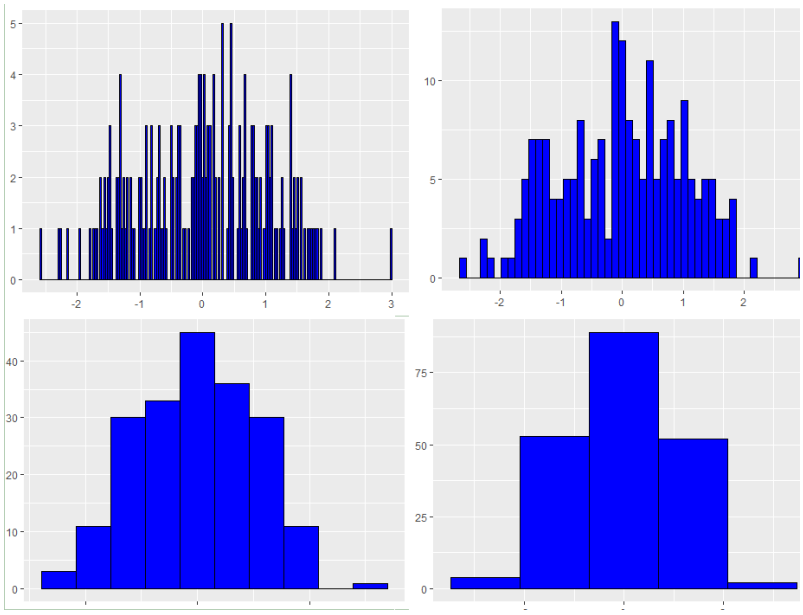
Другим вариантом, предоставляющим непрерывную оценку для функции плотности, является так называемая ядерная оценка плотности, задаваемая формулой

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

где h — некоторый положительный параметр, называемый шагом сглаживания, а K — ядро, то есть плотность некоторого распределения. При $n \to \infty$ величина $\widehat{f}_n(x)$ в каждой точке x сходится к

$$\frac{1}{h}\mathbf{E}K\left(\frac{x - X_1}{h}\right) = \frac{1}{h}\int_{\mathbb{R}} K(\frac{x-y}{h}) f_X(y)dy = \int_{\mathbb{R}} K(u) f_X(x - uh)du.$$

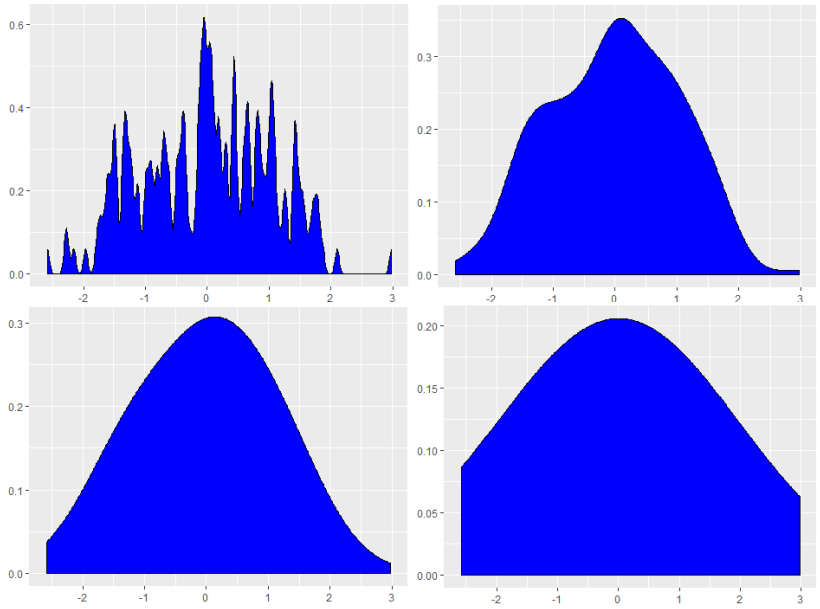
Рис. 2: Гистограммы для реализации стандартного нормального распределения с различным h



При $h \to 0$ данный интеграл сходится к $f_X(x)$.

Однако, данная оценка имеет те же проблемы, что и гистограмма — выбор h достаточно затруднителен из-за того, что слишком маленькое h делает плотность сконцентрированной в точках реализации, а большое "размазывает" функцию плотности, игнорируя ее особенности.

Рис. 3: Ядерные оценки плотности для реализации стандартного нормального распределения с различным h

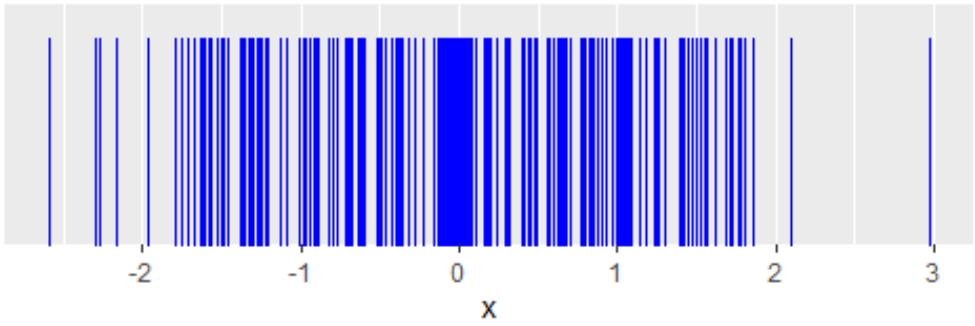


Отметим, что прежде чем оценивать плотность бывает полезным отразить наблюдения с помощью так называемого rug plot: одномерного графика, отражающего позиции точек.

1.4 Многомерные данные

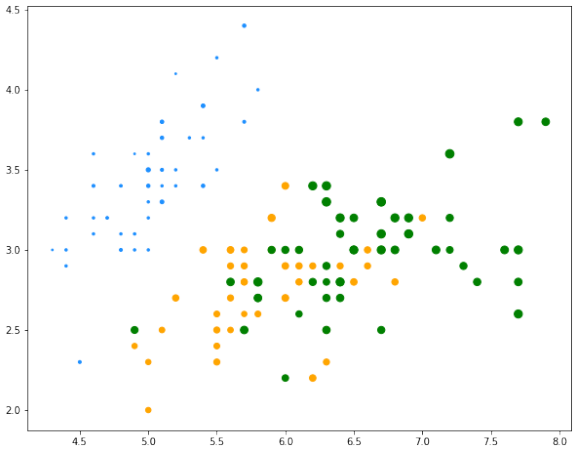
Для случая, когда элементы выборки представляют собой независимые одинаково распределенные векторы, мы используем другие подходы. В случае двумерных данных можно использовать диаграмму рассеяния scatterplot, отмечая точки в двух измерениях.

Рис. 4: Rug plot для реализации стандартного нормального распределения с различным h



Для данных размерности 3-4 мы можем отметить два параметра двумерной точкой, а остальные координаты задать, например, оттенком цвета, размером или формом. Этот подход наиболее удобен, если дополнительные переменные принимают небольшое количество значений или менее важны для нас.

Рис. 5: Диаграмма рассеяния для размеров лепестков ириса. Цвет отображает вид, а размер — размер чашечки цветка.



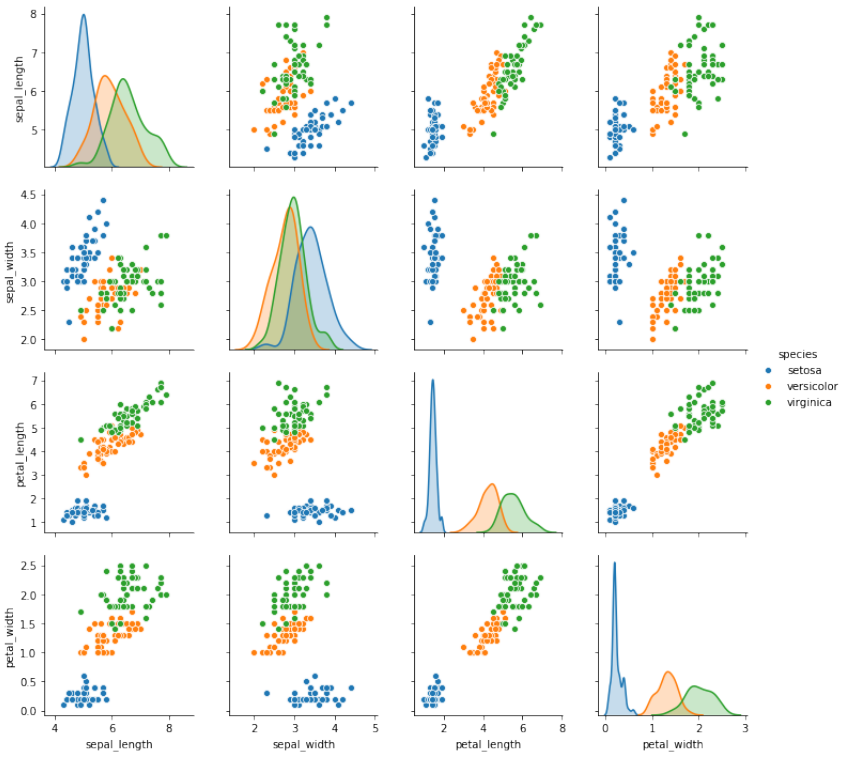
Также можно использовать матрицу диаграмм рассеяния scatterplot matrix. Это матрица из диаграмм рассеяния каждой из пар переменных (см. рисунок 6). К сожалению, даже если ни в одной из проекций данные не разделяются, при этом множества могут быть отделимы плоскостью. Одним из вариантов решения данной проблемы является так называемый тур — анимация, включающая перемещение из одной такого рода плоскости в другую. Мы не будем сталкиваться с такого рода визуализацией, но полезно иметь ее ввиду.

Если переменных, скажем, 6-8, то матрица диаграмм станет непомерно большой. В этом случае обычно стараются выбрать плоскую картинку (или хотя бы снизить размерность), похожую по расположению на наши трехмерные объекты. Об одном из таким подходов, называемых методом главных компонент, мы поговорим заметно позднее. А пока отметим метод RadViz, который переводит объемную картинку в плоскую

Если точек немного, а их размерность достаточно велика, то можно отображать каждую точку некоторым объектом, например, кривой, рассчитывая, что близость точек вызовет близость объектов и наоборот. Приведем четыре варианта:

- Параллельные кривые. В этом случае каждая точка $x_{i,1}, \ldots, x_{i,d}$ реализации отображается в виде ломаной с вершинами $(j, x_{i,j})$ (см. рисунок 7).
- Радарная диаграмма. Этот график близок к предыдущему, но ломаная располагается радиально, превращаясь в замкнутую (см. рисунок 8).

Рис. 6: Матрица диаграмм рассеяния для четырех параметров цветков ириса



- Кривые Эндрюса. В этом случае точка $(x_{i,1}, \ldots, x_{i,d})$ отображается кривой

$$f(t) = \frac{x_{i,1}}{\sqrt{2}} + x_{i,2} \sin t + x_{i,3} \cos t + x_{i,4} \sin 2t + \cdots$$

  Пример изображен на рисунке 9.

- Лица Чернова.

Существенно отличается от названных вариант, использующий в качестве изображений точек человеческие лица. Координаты точек при этом отражают различные параметры лица, например, овал, форму и размер глаз и так далее. С одной стороны, если объекты разделяются на несколько категорий, в каждой из которых точки достаточно близки, то глаз легко разделит лица на соответствующий категории. С другой стороны, координаты получаются существенно неравноправными (как, впрочем, и в прошлом методе). Скажем, овал лица бросается в глаза куда лучше чем форма ушей. Пример такого графика можно найти на рисунке 10.

1.5 Сравнение нескольких выборок

В случае если выборка $X_1, \ldots, X_n$ представляет собой объединение нескольких наборов $X_{1,1}, \ldots, X_{1,n_1}, \ldots, X_{k,1}, \ldots, X_{k,n_k}$, $n_1 + \cdots + n_k = n$, естественным является желание сравнить эти наборы друг с другом. При небольшом k мы можем построить столбцовые диаграммы, гистограммы или ядерные оценки плотности на одном графике.

Однако, если k достаточно велико, визуально будет тяжело сравнить столько графиков, поэтому в таких случаях используют другие подходы. Одним из вариантов является скрипичный график (violin plot), представляющий собой вертикально расположенные симметризованные графики плотностей. Пример такого графика вы найдете на рисунке 11.

Однако, и этот вариант тяжело читается при достаточно большом k, из-за чего информацию, содержащуюся в каждом графике убирают до описанных в начале лекции квартилей и медианы. Для этого используется так называемая диаграмма размаха (или как ее называют в народе "ящик с усами"). Пример такого графика вы найдете на рисунке 12. Существует ряд других близких форм визуализации: box percentile plot, vaseplot и так далее, на которых мы останавливаться подробнее не будем.

Рис. 7: Цветы ириса в параллельных координатах

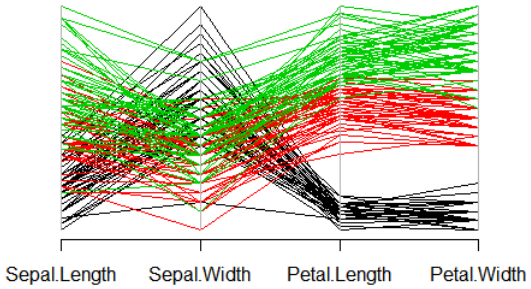


Рис. 8: Цветы ириса в радарной диаграмме (виды представлены одной ломаной)

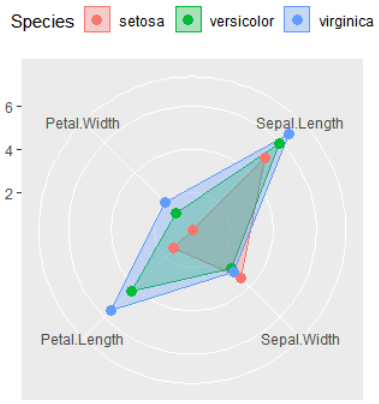


Рис. 9: Цветы ириса в виде кривой Эндрюса

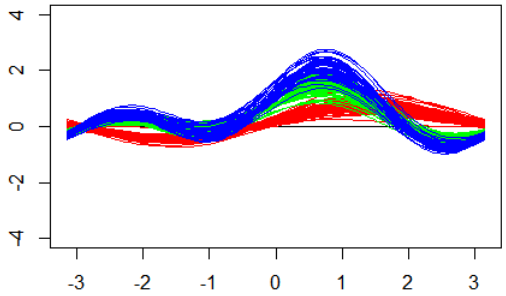


Рис. 10: 32 автомобиля в лицах Чернова: высота лица и высота носа характеризует расход топлива, ширина — число цилиндров, высота шевелюры — число передач

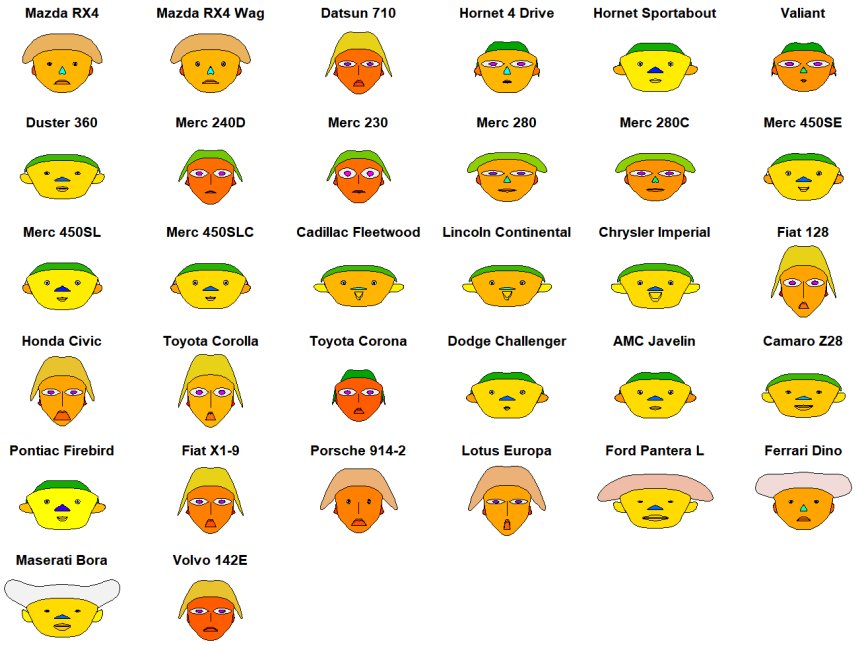


Рис. 11: Скрипичный график для болевых порогов людей с четырьмя цветами волос

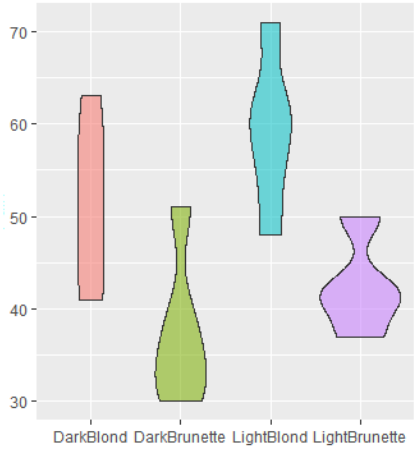


Рис. 12: Диаграммы размаха для очков, набираемых игроками в 10 регбийных играх сезона