Описательные статистики

Сегодня обсудим интересную базу данных – базу с экспериментальными данными по шоколадным тортам.

Описание базы данных.

Загрузим базу данных по ссылке:

```
cakes <- read.csv("cake.csv")
```

Посмотрим на неё:

```
View(cakes)
```

Для вывода описательных статистик в R есть специальная функция summary():

```
summary(cakes)
```

```
##        X            replicate  recipe  temperature      angle
##  Min.   :  1.00   Min.   : 1   A:90   Min.   :175   Min.   :18.00
##  1st Qu.: 68.25   1st Qu.: 4   B:90   1st Qu.:185   1st Qu.:26.00
##  Median :135.50   Median : 8   C:90   Median :200   Median :31.00
##  Mean   :135.50   Mean   : 8          Mean   :200   Mean   :32.12
##  3rd Qu.:202.75   3rd Qu.:12          3rd Qu.:215   3rd Qu.:36.75
##  Max.   :270.00   Max.   :15          Max.   :225   Max.   :63.00
##       temp
##  Min.   :175
##  1st Qu.:185
##  Median :200
##  Mean   :200
##  3rd Qu.:215
##  Max.   :225
```

Для количественных переменных эта функция выдает минимальное и максимальное значение, среднее арифметическое, медиану, нижний (`1st Qu.`) и верхний (`3rd Qu.`) квартиль. Нижний квартиль – значение, которое 25% значений в выборке не превышают, а верхний квартиль – значение, которое 75% значений в выборке не превышают. Для качественных переменных (текстовые, факторные), R будет выводить количество значений по каждой группе (уровню).

В данном случае по выдаче R мы можем определить следующее. Всего в базе данных у нас 270 наблюдений (переменная X здесь служит id наблюдений, а её максимальное значение 270), значит, в рамках исследования было приготовлено 270 шоколадных тортов. Минимальная температура, при которой выпекали торты, равна 175 градусам, максимальная — 225. Средняя температура, при которой выпекали торты, равна 200. Медианное значение температуры в данном случае совпадает со средним значением — в половине случаев температура при выпечке не превышала 200 градусов. Нижний квартиль равен 185 градусам — в 25% случаев торты выпекались при температуре не выше 185 градусов, верхний квартиль равен 215 — 75% случаев температура не превышала 215 градусов (или в 25% случаев превышала!).

Необязательно выводить описательные статистики для всех переменных в базе данных, можно вывести описание одной переменной:

```
summary(cakes$temperature)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     175     185     200     200     215     225
```

Или проделать то же самое для нескольких переменных:

```
summary(cakes[5:6])  # в скобках - индексы столбцов
##      angle            temp
##  Min.   :18.00   Min.   :175
##  1st Qu.:26.00   1st Qu.:185
##  Median :31.00   Median :200
##  Mean   :32.12   Mean   :200
##  3rd Qu.:36.75   3rd Qu.:215
##  Max.   :63.00   Max.   :225
```

Или так, по названиям столбцов:

```
summary(cakes[, c("angle", "temp")])
##      angle            temp
##  Min.   :18.00   Min.   :175
##  1st Qu.:26.00   1st Qu.:185
##  Median :31.00   Median :200
##  Mean   :32.12   Mean   :200
##  3rd Qu.:36.75   3rd Qu.:215
##  Max.   :63.00   Max.   :225
```

Точно так же необязательно выводить все статистики сразу. Можно запрашивать по отдельности:

```
min(cakes$temp)  # минимум
## [1] 175
max(cakes$temp)  # максимум
## [1] 225
mean(cakes$temp)  # среднее
## [1] 200
median(cakes$temp)  # медиана
## [1] 200
```

Описательные статистики можно находить не только для всей переменной, но и для тех значений, которые соответствуют какому-то условию.

```
mean(cakes$temp[cakes $recipe == 'A'])
```

Или нескольким условиям:

```
mean(cakes$temp[cakes$recipe == "A" & cakes$replicate == 8])

sd(cakes$temp[cakes$recipe != "A" & cakes$angle > 35])
```

Теперь посмотрим на квантили. Для примера запросим квантиль уровня 0.25 для переменной `temp`, то есть значение температуры, которое 25% значений в выборке не превышают.

```
quantile(cakes$temp, 0.25)  # переменная, а затем уровень квантиля
## 25%
## 185
```

Выдача для квантиля выглядит интересно: помимо самого значения выводится уровень 25%. Результат, который возвращает функция `quantile()`, является поименованным вектором, у каждого элемента вектора есть название. Здесь элемент один, и название одно, `25%`.

Если мы запросим сразу несколько квантилей, перечислим уровни в виде вектора, то всё тоже сработает:

```
quantile(cakes$temp, c(0.25, 0.5, 0.75))
## 25% 50% 75%
## 185 200 215
```

Можем задавать уровни в виде последовательности с заданным шагом. Вызовем децили — квантили с уровнями, кратными 10:

```
quantile(cakes$temp, seq(from=0, to=1, by=0.1))
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  175  175  185  185  195  200  205  215  215  225  225
```

Ещё есть функция `fivenum()`, которая возвращает описательные статистики по Тьюки, те, которые используются для построения ящика с усами: минимум, нижняя граница типичных значений, медиана, верхняя граница типичных значений, максимум.

```
fivenum(cakes$temp)
## [1] 175 185 200 215 225
```

Это не единственные описательные статистики, которые можно вывести. Часто нас интересует не только среднее (или медианное) значение, а разброс значений относительно этого среднего. Для этого можем посчитать *выборочную дисперсию* или *стандартное отклонение*.

```
var(cakes$temp)  # дисперсия
## [1] 292.7509
sd(cakes$temp)  # стандартное отклонение
## [1] 17.10997
```

Но сами по себе эти значения не очень информативны – по ним сложно понять, насколько однородны наши данные (сильно ли они разбросаны относительно среднего значения). Для того, чтобы оценить степень однородности наших данных, нашей выборки, можно воспользоваться

таким показателем как *коэффициент вариации*. Коэффициент вариации считается несложно: стандартное отклонение нужно поделить на среднее значение. Обычно значение коэффициента вариации, взятое по модулю, лежит в пределах от 0 до 1, но иногда, если данные очень разнородны (стандартное отклонение большое), оно может быть больше 1.

Часто коэффициент вариации выражают в процентах. Давайте напишем код, который будет считать коэффициент вариации в процентах для переменной `angle`.

```
sd(cakes$temp)/mean(cakes$temp) * 100
```

```
## [1] 8.554983
```

В данном случае все показатели считаются без проблем, потому что в базе данных все строки полностью заполнены. Если среди значений встречаются пропущенные (`NA`), то и результат тоже будет `NA`. Чтобы решить эту проблему, нужно прописать дополнительный аргумент `na.rm = TRUE`, который говорит R не учитывать пропущенные значения при расчете статистик (из самой базы значения при этом не выкидываются!).

```
ages <- c(23, 25, 27, NA)
mean(ages, na.rm = TRUE)
```

```
## [1] 25
```

Пока мы обсудили только описательные статистики для количественных переменных. А как быть с качественными? Какую информацию по ним можно получить? Число наблюдений, соответствующих каждому значению (классу):

```
table(cakes$recipe)
```

```
##
##  A  B  C
## 90 90 90
```

В прошлый раз мы говорили о том, что для показателей, измеренных в качественной шкале, вычислять среднее или медиану бессмысленно, нужно смотреть на моду. Искать специальную функцию не нужно, достаточно помнить, что мода — это значение, которое встречается в выборке чаще всего (да, мода может быть не одна).

Кстати, для количественных переменных функция `table()` тоже хорошо работает:

```
table(cakes$angle)
```

```
##
## 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
##  1  2  3 10  5  8 18 13 14 13 21 12 13 14  8 19  4 21  3 11  4  6  5  3  5
## 43 44 45 46 47 48 49 51 52 53 55 57 58 61 63
##  6  1  4  7  6  1  1  1  1  1  1  1  1  1  1
```

В заключение первой части знакомства с описательными статистиками в R давайте установим библиотеку `psych`, которая используется для психометрических исследований и посмотрим, какие у неё есть возможности, связанные с описанием данных.

```
install.packages("psych")
```

Обратимся к ней:

```
library(psych)
```

Теперь запросим описательные статистики с помощью функции `describe()`:

```
describe(cakes$temp)
```

```
##    vars   n mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 270  200 17.11    200     200 22.24 175 225    50    0    -1.28
##      se
## X1 1.04
```

Что есть что?

- `vars`: число описываемых переменных (не путать с `var` для дисперсии);
- `n`: число наблюдений;
- `mean`: среднее арифметическое, выборочное среднее;
- `sd`: стандартное отклонение;
- `median`: медиана;
- `trimmed`: усечённое среднее, среднее по цензурированной выборке (см. ниже);
- `mad`: медианное значение абсолютного отклонения от медианы (нам не понадобится);
- `min`, `max`: минимальное и максимальное значение;
- `range`: размах;
- `skew`: коэффициент асимметрии или скошенности (см.нижк);
- `kurtosis`: коэффициент эксцесса (см. ниже);
- `se`: стандартная ошибка среднего;

Подробнее про некоторые статистики.

Усечённое среднее, среднее по цензурированной выборке

- Считается так: выборка упорядочивается по возрастанию, из неё убирается 5% наблюдений слева и справа (наименьшие и наибольшие), потом по такой усечённой или цензурированной выборке считается обычное среднее арифметическое.
- Наравне с медианой считается более устойчивой оценкой среднего, так как после усечения выборки такой показатель уже несильно зависит от слишком больших или слишном маленьких (нетипичных) значений в выборке. То есть, при наличии нетипичных наблюдений в выборке (выбросов) такое среднее более адекватно отражает реальность, чем обычное среднее арифметическое.

Коэффициент асимметрии

- Показатель принимает значения примерно от -3 до 3. Значение 0 соответствует симметричному распределению (например, нормальному, вспомните график плотности, симметричный относительно математического ожидания). Значения меньше 0 соответствуют распределению, которое скошено влево (длинный хвост «слева»), значения больше 0 соответствуют распределению, которое скошено вправо (длинный «хвост» справа).
- В нашем случае распределение почти симметричное, коэффициент близок к нулю, но при это оно немного скошено вправо, поэтому значение больше 0.

Коэффициент эксцесса

- Показатель принимает значения примерно от -3 до 3 и отвечает за выраженность пика распределения. Чем больше значение коэффициента, тем более выраженный пик. Стандартное нормальное распределение имеет коэффициент эксцесса равный 0. Отрицательные значения коэффициента соответствуют более «плоским» и «гладким» распределениям, у которых пик не такой заметный.

- В нашем случае распределение несильно отличается от нормального, поэтому коэффициент близок к нулю.

Библиотека `psych` удобна тем, что она содержит функцию `describeBy()`, которая позволяет выводить описательные статистики по группам. Нет необходимости отфильтровывать нужные строки и сохранять их в отдельные датасеты, можно просто указать группирующую переменную. Выведем описательные статистики для переменной `temp` отдельно для каждого рецепта (здесь у всех групп всё будет одинаково в силу специфики экспериментальных данных):

```
describeBy(cakes$temp, cakes$recipe)
```

```
##
##  Descriptive statistics by group
## group: A
##    vars  n mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 90  200 17.17    200     200 22.24 175 225    50    0    -1.31
##      se
## X1 1.81
## ------------------------------------------------------------
## group: B
##    vars  n mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 90  200 17.17    200     200 22.24 175 225    50    0    -1.31
##      se
## X1 1.81
## ------------------------------------------------------------
## group: C
##    vars  n mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 90  200 17.17    200     200 22.24 175 225    50    0    -1.31
##      se
## X1 1.81
```