Типичное значение выборки. Среднее или медиана.
Вариант 15.07.2004

Анализируем население городов России в 1959 году. Данные получены по результатам переписи населения.

Описание данных.
Население в тысячах человек. В выборку попали населенные пункты, которые считаются городами России в настоящее время, даже если в 1959 году они еще не считались городами.
Данные находятся в файле town_1959_2.csv

```
# Задаем рабочую папку и начинаем анализ с импорта данных в R.

setwd("C:/aaa/_R/_R_лекции/_lectures_2012/R_занятие_3")

town.1959 <- read.table("town_1959_2.csv", header=T, sep="," )

# Посмотрим на данные.
# Зачем смотреть, все вроде бы правильно?
#  Например, если бы мы пропустили любой из параметров
# header=T или sep=",", результат импорта был бы неправильным.

town.1959

#   Посмотрим описательные статистики, характеризующие выборку.
summary(town.1959[,3])
#     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#     0.10   10.70   19.25   52.93   37.97 5046.00


#   Наблюдение 1. Среднее арифметическое больше 3 квартили!
#   Уточним.

sum(town.1959[,3] < 52.93)/nrow(town.1959) * 100
#   [1] 82.37052

#   Другой способ посчитать то же самое
#   mean(town.1959[,3]<mean(town.1959[,3])) * 100


#   Наблюдение 2. Если в качестве населения типичного города
#   России взять среднее арифметическое, то 82% городов России
#   имеет население меньше, чем население типичного города.
#   Что вызывает дискомфорт. Такое наблюдение не воспринимается
как типичное...

# Сколько всего наблюдений?
nrow(town.1959)
#   [1] 1004

(52.93 -  45.00)/ 52.93 * 100
```

```
#   [1] 14.98205

2/ 1004 * 100

#   [1] 0.1992032

#   Наблюдение 3. После отбрасывания 0.2% наблюдений среднее
#   арифметическое уменьшилось на 15%
#   Медиана же уменьшилась на 100 человек

#   Если признать, что Москва и Санкт-Петербург выбросы, и
#   исключить их из выборки, получим следующее

summary(town.1959[-c(1, 2),3])
#     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#     0.10   10.70   19.15   45.00   37.55  941.00
```

Вывод.
Если выборка содержит выбросы, т. е. аномально большие или аномально маленькие наблюдения, то вычисление среднего арифметического становится ненадежным методом определения типичного значения.

Медиана лучше.

Некоторые полезные команды.

Сосчитать только среднее

```
mean(town.1959[ , 3])
```

Сосчитать только медиану

```
median(town.1959[,3])
```

Сосчитать усеченное среднее, p=0.95

```
mean(town.1959[ , 3], trim = 0.025)
```

Допустимо использовать усеченное среднее, но не моду.
Но! Усеченное среднее плохо воспринимается заказчиком.

Сколько выбросов и каково распределение данных?

На гистограмме видны только выбросы

```
hist(town.1959[ , 3])
```

Но сколько их?

На гистограмме

```
hist(log(town.1959[ , 3]), breaks = 44)
```

видно, что у нас 3 выброса

Внешне распределение похоже на лог-нормальное.
(Вопрос по не пройденному материалу: почему все же не логнормальное?)