Санкт-Петербургский Государственный Политехнический Университет

Расчетное задание

по курсу: «Математическая статистика»
студент: Зенцев Ф.К., гр. 3057/2
преподаватель: Галинина О.С.

Оглавление

Выборка

В задании анализируется цена на импорт сырой нефти в США. В качестве выборки были выбраны логарифмические приращения значений цены за одну единицу сырой нефти в долларах за месяц с 1973 по 2009 года.

Удаление из выборки «выбросов» проводилось следующим образом: из соответствующего вариационного ряда удалялись несколько значений с обоих концов.

Код:

```
main_sample <<- sort(main_sample)

for(i in 1:clipCount) main_sample <<- main_sample[-1]
for(i in 1:clipCount) main_sample <<-
                      main_sample[-length(main_sample)]
```

Далее по отчету результаты приведены для `clipCount` равного 50. Выводы из полученных результатов в отдельный пункт не выделены – делаются внутри остальных разделов.

Детали реализации

Задание выполнено на специальном языке статистического анализа - языке программирования R.

Подсчет параметров

Выборочные описательные статистики

Цель описательной (дескриптивной) статистики - обработка эмпирических данных, их систематизация, наглядное представление в форме графиков и таблиц, а также их количественное описание посредством основных статистических показателей.

Показатели центра распределения

Показателями центра распределения обычно называют среднее, моду и медиану. Эти показатели используются для определения наиболее типичных значений совокупности.

Код:

```
mean(sample)

median(sample)

as.numeric(names(sort(-table(sample)))[1])
```

Результат:

```
Mean:

-0.003983287

Median:

-0.00322321

Mode of sample:

0
```

Пояснения:

Мода выборки это элемент, который повторяется чаще всего в выборке. Ясно, что такое определение не походит в данном случае, так как выборка не состоит из целых чисел и соответственно не найдется такого элемента, который повторялся бы дважды. В таком случае всю выборку разбивают на интервалы (см. гистограмма) и, далее в качестве моды выбирается то значение из интервалов, где гистограмма достигает своего наибольшего значения. Далее в отчете можно будет убедиться в правильном ответе 0 для данной выборки.

Показатели вариации

Среднеквадратичное отклонение и коэффициент вариации. Вычисление абсолютного (среднеквадратичное отклонение) и относительного (коэффициент вариации) показателей вариации.

Код:

```
sd(sample)

sd(sample) / mean(sample)
```

Результат:

```
Standard deviation:

0.04115969

Coefficient of variation:

-10.33310
```

Показатели формы распределения

Ассиметрия и эксцесс.

Код:

```
skewness(sample)

kurtosis(sample)
```

Результат:

```
Skewness:

0.23591

Kurtosis:

3.016410
```

Пояснения:

Неформально говоря, коэффициент асимметрии показывает насколько несимметрична выборка относительного своего среднего значения. Понятно, что в случае «хорошей» симметрии этот коэффициент должен быть близок к нулю. Далее приведенные гистограммы позволят убедиться в правильности полученных здесь результатов. Эксцесс же позволяет судить об остроте пика гистограммы – положителен в случае острого пика.

Гистограмма

Для выбора числа интервалов гистограммы были опробованы правила Фридмана-Диакониса

$$size = 2 * IQR(X) * n^{-1/3}$$

где IQR – интерквартальное среднее выборки, а n – объем выборки

и Стерджесса

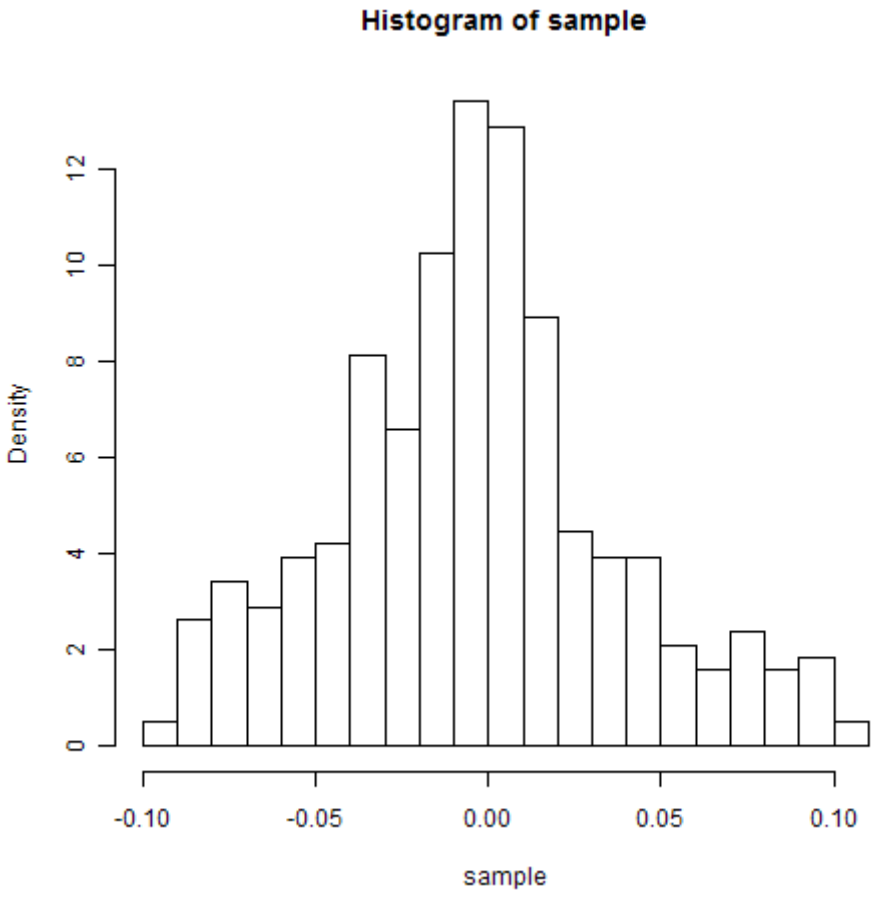
$$k = \lceil 1 + \log_2 n \rceil$$

Далее результаты приведены для правила Фридмана-Диакониса.

Код:

```
# Use Freedman-Diaconis rule (2 * IQR(x) * n^(-1/3)) to select
number of histogram bins
my_hist <<- hist(sample, breaks = "FD", plot = TRUE, freq =
FALSE)
```

Результат:



Аппроксимация нормальным распределением

Полагаем, что искомая выборка получена из нормального распределения. Далее находим с помощью метода моментов параметры такого нормального распределения. После этого генерируем выборку из такого «теоретического» распределения и считаем все те параметры, которые обсуждались до этого момента, строим гистограмму.

Код:

```
# Apply method of moments,
# assuming that sample is from normal distribution
m1 = moment(main_sample, order = 1)
m2 = moment(main_sample, order = 2)

main_sample_sigma <<- sqrt(m2 - m1 * m1)
main_sample_mu <<- m1

# Generate sample from normal distribution (main_sample_mu,
main_sample_sigma)

theor_sample <<- rnorm(length(main_sample), sd =
main_sample_sigma, mean = main_sample_mu)
```

Результат:

см. Приложение

Пояснения:

Незначительные различия в значениях параметров и «похожесть» гистограммы позволяет нам высказать догадку о том, что искомое распределение как раз и является нормальным. В дальнейшем (см. проверка гипотез) этот вопрос будет обсуждаться более строго.

Правило «трех сигм»

Это правило еще называют «правило 68-95-99.7» - наглядно показывает, что это правило означает. В отрезок $[\mu-3\sigma;\mu+3\sigma]$ попадает ~99.7% значений выборки из нормального распределения с параметрами μ,σ

Соответствующие проценты из названия правила для отрезков $\mu\pm2\sigma,\mu\pm\sigma$

Код:

```
b1 = 0
b2 = 0
b3 = 0
mu = mean(sample)
sigma = sd(sample)

for(i in 1:length(sample))
{
  if(sample[i] < mu + sigma && sample[i] > mu - sigma) {
      b1 = b1 + 1;
     }

  if(sample[i] < mu + 2 * sigma && sample[i] > mu - 2 * sigma) {
      b2 = b2 + 1;
    }

  if(sample[i] < mu + 3 * sigma && sample[i] > mu - 3 * sigma) {
      b3 = b3 + 1;
    }
}

cat(b1 / length(sample) * 100)
cat(b2 / length(sample) * 100)
cat(b3 / length(sample) * 100)
```

Результат:

```
68.50394%
92.91339%
100%
```

Пояснения:

Этот результат также подтверждает высказанную ранее догадку о виде распределении.

Проверка гипотез

Наиболее удобным в нашем случае способом представления результатов будет вывод достигаемого уровня значимости (p-value) – наименьшей величины уровня значимости, при котором нулевая гипотеза отвергается для данного значения статистики критерия.

Как правило, в практических задачах нет никакого разумного правила для выбора фиксированного уровня значимости. Выбирая метод достигаемого уровня значимости, мы можем сделать процедуру принятия решения более гибкой - чем меньшее значение p-value мы наблюдаем, тем сильнее свидетельствует совокупность наблюдений против нулевой гипотезы. Вообще говоря, часто нулевую гипотезу отвергают при значениях p-value меньших 0.01 или 0.05.

Гипотеза о виде распределения

Формулировка гипотезы:

$H_0 = \{F_n(x) = F_{N(\mu,\sigma)}(x)\}$, где μ, σ - параметры найденные ранее.

Критерий Колмогорова

Статистика критерия:

$$T = \sqrt{n} \sup_x | F_n(x) - F_{N(\mu,\sigma)}(x) |$$

Критическая область:

$T \geq \lambda_\alpha$, где λ_α это квантиль распределения Колмогорова.

Код:

```
test_result = ks.test(sample, "pnorm", mean = main_sample_mu, sd
= main_sample_sigma)
cat(test_result$method)
cat(test_result$p.value)
```

Результат:

```
One-sample Kolmogorov-Smirnov test

p-value:
0.3304615
```

Критерий Пирсона

Статистика критерия:

$$T = \sum_{j=1}^{N} \frac{(\upsilon_j - np_j^0)^2}{np_j^0}$$

Критическая область:

$T \geq t_\alpha$, где t_α это $(1-\alpha)$-квантиль распределения хи-квадрат с $(N-1)$ степенями свободы.

Код:

```
test_result = pearson.test(sample)
cat(test_result$method)
cat(test_result$p.value)
```

Результат:

```
Pearson chi-square normality test

p-value:
0.2463146
```

Пояснения:

Как видно из полученных результатов нулевая гипотеза не отвергается для «хорошего» уровня значимости, что уже в более строгом смысле подтверждает нашу догадку о распределении.

Гипотеза однородности

Проверки гипотез однородности и независимости предполагают наличие двух выборок. Поделим исходную выборку пополам.

Формулировка гипотезы:

$$H_0 = \{F_1(x) = F_2(x)\}$$

Критерий Колмогорова-Смирнова

Статистика критерия:

$$T = \sqrt{\frac{n+m}{nm}} \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

Критическая область:

$T \geq \lambda_\alpha$, где λ_α это квантиль распределения Колмогорова.

Код:

```
test_result = ks.test(sample1, sample2)
cat(test_result$method)
cat(test_result$p.value)
```

Результат:

```
Two-sample Kolmogorov-Smirnov test

p-value:
0.1195520
```

Критерий Уилкоксона-Манна-Уитни

Статистика критерия:

$T = \sum_{i=1}^{n} R_i$, где R_i - ранги элементов первой выборки в объединенном вариационном ряду.

Критические области:

$$T \le \frac{nm}{2} - \sqrt{\frac{nm(n+m+1)}{12}} t_\alpha$$

$$| T - \frac{nm}{2} | \ge \sqrt{\frac{nm(n+m+1)}{12}} t_{\alpha/2}$$

где t_α- квантили нормального распределения $N(\frac{nm}{2}, \sqrt{\frac{nm(n+m+1)}{12}})$

Код:

```
test_result = wilcox.test(sample1, sample2)
cat(test_result$method)
cat(test_result$p.value)
```

Результат:

```
Wilcoxon rank sum test with continuity correction

p-value:
0.2670209
```

Критерии Фишера и Стьюдента

Эти критерии проверяют гипотезу однородности в менее общем виде, чем ранее. Гипотеза формулируется в предположении, что обе исходные выборки получены из некоторых нормальных распределений:

$$X = (X_1, ..., X_n) \sim N(\mu_1, \sigma_1)$$
$$Y = (Y_1, ..., Y_m) \sim N(\mu_2, \sigma_2)$$

Формулировки гипотез:

$$H_0^f = \{\sigma_1^2 = \sigma_2^2\}, H_0^t = \{\mu_1 = \mu_2\}$$

Статистики критериев:

$$T^f = \frac{S^2(X)}{S^2(Y)}, T^t = \sqrt{\frac{nm}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)S^2(X) + (m-1)S^2(Y)}{n+m-2}}}$$

Критические области:

$f_{\alpha/2} \le T^f \le f_{1-\alpha/2}, |T^t| \ge \tau_{1-\alpha/2, n+m-2}$, где f_α- квантили распределения Фишера, а τ_α - квантили распределения Стьюдента.

Код:

```
test_result = var.test(sample1, sample2)
cat(test_result$method)
cat(test_result$p.value)

test_result = t.test(sample1, sample1, var.equal = TRUE)
cat(test_result$method)
cat(test_result$p.value)
```

Результат:

```
Wilcoxon rank sum test with continuity correction

p-value:
0.2670209

F test to compare two variances

p-value:
0.08649648
```

Гипотеза независимости

Формулировка гипотезы:

$H_0 = \{X \text{ и } Y \text{ независимы } \}$

Критерий хи-квадрат

Статистика критерия:

$$T = \sum_{i,j} \frac{(\upsilon_{ij} - np_i^x p_j^y)^2}{np_i^x p_j^y}$$

Критическая область:

$T \geq t_\alpha$, где t_α это $(1-\alpha)$-квантиль распределения хи-квадрат с $(k-1)(m-1)$ степенями свободы.

Код:

```
test_result = chisq.test(sample1, sample2)

cat(test_result$method)
cat(test_result$p.value)
```

Результат:

```
Pearson's Chi-squared test

p-value:
0.2407977
```

Приложение

Параметры для выборки из теоретического распределения

Mean:

-0.004616894

Median:

-0.007078231

Mode of sample:

-0.1083534

Standard deviation:

0.03648177

Coefficient of variation:

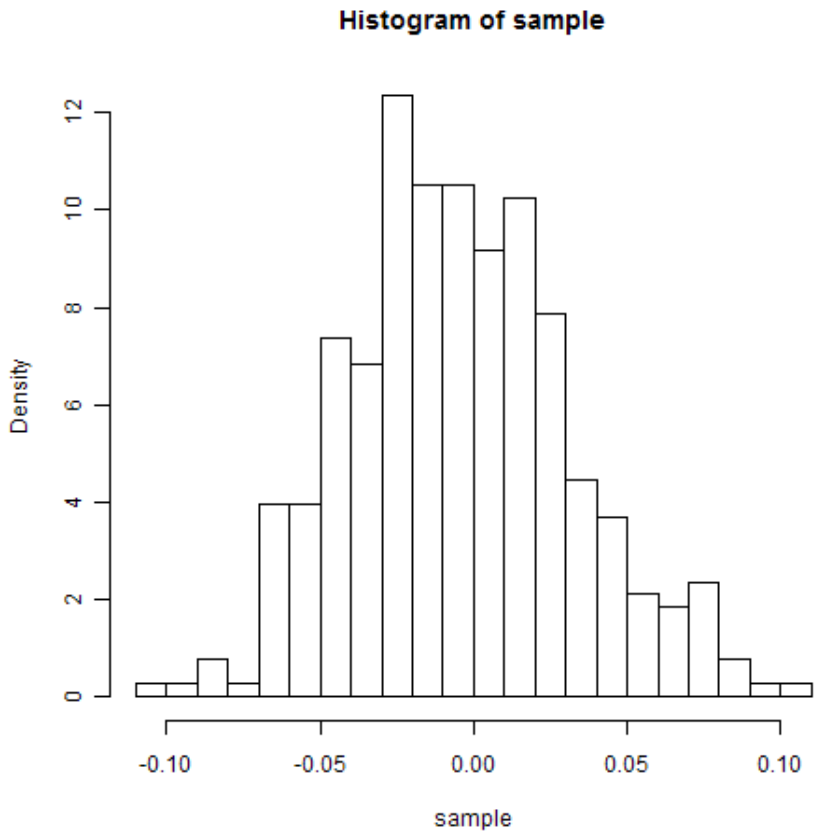
-7.9018

Skewness:

0.2593295

Kurtosis:

2.976189



Код программы

Результаты

Использованная литература

1. Н.И.Чернова «Математическая статистика. Учебное пособие», Новосибирск 2007
2. Ивченко Г.И., Медведев Ю.И. «Математическая статистика: Учеб. пособие для втузов» М.: Высш. школа, 1984
3. Конспект лекций и упражнений по курсу «Математическая статистика» (Преподаватели: Мясникова Екатерина Марковна и Галинина Ольга Сергеевна)
4. An R Introduction to Statistics: http://www.r-tutor.com/