ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Отчет по лабораторной работе № 3   
«Точечное оценивание параметров распределения. Критерий согласия Пирсона»

наименование лабораторной работы

Выполнил: студент гр. 1ФКЦ

Ахметов Руслан Олегович

Проверил: кандидат технических наук, доцент

Прудников Вадим Борисович

Уфа – 2023

1.В предположении логнормального распределения для среднемесячной заработной платы рассчитаем точечные оценки параметров распределения. Для выборки отбираем только ненулевые (положительные) значения заработной платы:   
var3.poswage <- var3[var3[,'wage']>0,]$wage.

poswage (от англ. positive wage)

Результаты оптимизации по методу максимального правдоподобия сохраняем в списке mle.list:  
mle.list <- fitdist(var3.poswage, distr = 'lnorm', method = 'mle').

method="mle" (метод максимального правдоподобия)

method="mme" (метод моментов)

method = "qme" (метод квантилей)

Точечные оценки для мат.ожидания и стандартного отклонения mle.list и sigma.hat соответственно равны 9.8 и 0.651:

m.hat <- mle.list$estimate[1],

sigma.hat <- mle.list$estimate[2].



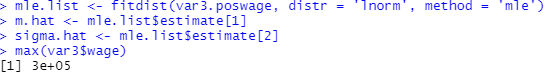


Рисунок 1.



Рисунок 2.

2.Построим гистограмму относительных частот. Для этого определим границы изменения заработной платы в виде от 0 до 250 000 руб. (максимальный размер заработной платы в рассматриваемой выборке). Разобьем его на 51 интервал и построим гистограмму:

x.arg <- seq(0, 250000, length =51);

h <- hist (var3.poswage, breaks =x.arg, freq=F, col = 'grey90');

lines(x.arg, dlnorm(x.arg, m.hat, sigma.hat), lty =1, col = 'blue', lwd = 2).





Рисунок 3.

Функция lines позволила наложить на гистограмму график плотности теоретического распределения с параметрами, равными найденным точечным оценкам. Результат представлен на рис. 4.

Функция lines() рисует линию текущим цветом рисования.   
Функция lines() в R используется для добавления линий различных типов, цветов и ширины к существующему графику.   
Синтаксис: lines (x, y, col, lwd, lty)

Видно, что график теоретической плотности достаточно хорошо описывает гистограмму относительных частот. При выполнении задали также необходимые наименования графика и названия осей.

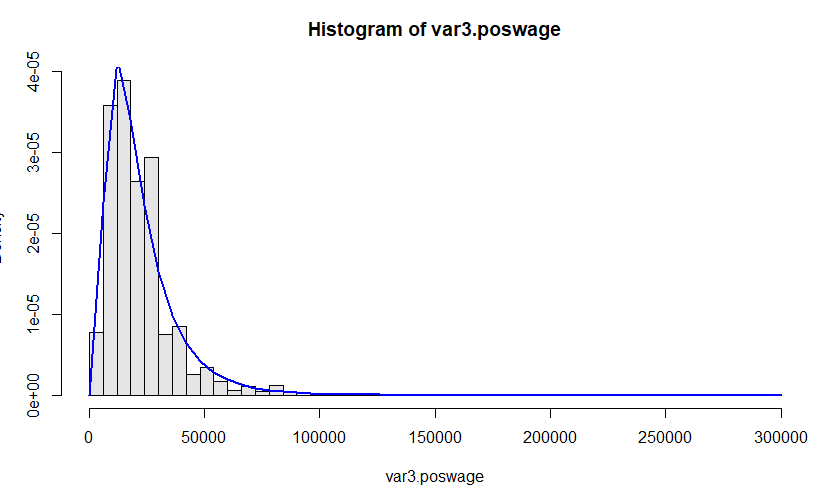


Рисунок 4 – Гистограмма среднемесячной заработной платы с наложением теоретической функции плотности

3.Проверим соответствие выборочного распределения логнормальному с помощью теста Колмогорова-Смирнова.

Нулевая гипотеза H0 состоит в том, что выборочное распределение соответствует предполагаемому. Альтернативная гипотеза H1 заключается в том, что такого соответствия нет.

Тестируемое распределение задается параметром функции (в нашем случае ‘plnorm’), также задаются математическое ожидание и стандартное отклонение, в качестве которых мы используем ранее найденные выборочные характеристики m.hat, sigma.hat:

ks.test(var3.poswage, plnorm, m.hat, sigma.hat)

Результат проведения теста представлен на рис. 5.

Как и в случае нормального распределения, p-value (вероятность того, что выборки получены из распределений с равными математическими ожиданиями (при верной H0) значительно меньше любого разумного уровня значимости, поэтому H0 отвергается, соответствия логнормальному распределению также нет.

Однако, необходимо обратить внимание, что в данном случае p-value на 8 порядков больше, чем в случае проверки гипотезы о соответствии нормальному распределению. Из двух указанных распределений логнормальное лучше описывает рассматриваемую выборку.

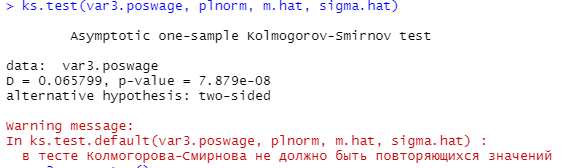


Рисунок 5 – Результаты теста Колмогорова-Смирнова

4.Теперь проверим соответствие выборочного распределения логнормальному на основе хиквадрат критерия согласия Пирсона.

Для подсчета фактических частот воспользуемся атрибутом counts построенной гистограммы относительных частот h: ccc <- h$counts. Поскольку для значений, начиная с 22-го интервала число значений мало, соответствующие частоты подлежат объединению:

ccc2 <- vector()

ccc2[1:21] <- ccc[1:21];

ccc2[22] <- sum(ccc[22:50]); ccc2

b <- h$breaks

b <- b[1:22]

sum.n <- sum(ccc2)

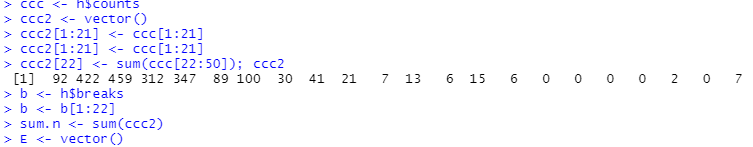
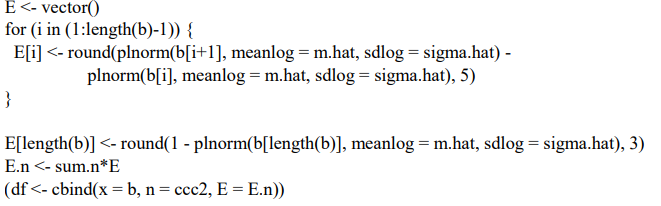


Рисунок 6.

Расчет теоретических частот несколько сложнее. Расчет вероятностей принятия значения из интервала производится в цикле по известной формуле теории вероятностей 𝑃(𝑎 ≤ 𝑋 ≤ 𝑏) = 𝐹(𝑏) − 𝐹(𝑎). При этом использована функция логнормального распределения, задаваемая командой plnorm.

Вариант исходного кода приведен ниже.



В результате набор данных df содержит как фактические, так и теоретические частоты (рис. 7). Напомним, в статистике любая сумма квадратов отклонений связана с числом степеней свободы df (degress of freedom), которое показывает, сколько независимых отклонений из n возможных участвует в образовании данной суммы квадратов.

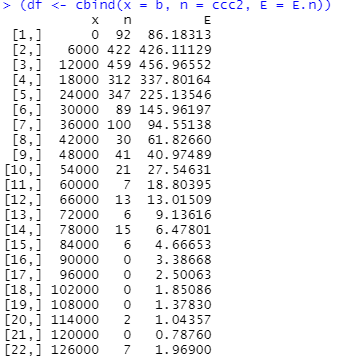


Рисунок 7 – Набор данных, содержащий фактические (n) и теоретические (E) частоты

Рассчитаем значение тестовой статистики Пирсона: chi2 <- sum((df[,2]-df[,3])^2/df[,3]); оно равно 82.4507694346386



Напомним, что нулевая гипотеза H0 состоит в том, что выборка получена из нормального распределения. При справедливости нулевой гипотезы тестовая статистика имеет распределение хи-квадрат с k-d-1 степенями свободы, где d – число параметров теоретического распределения, которые были оценены на основании выборки.

В нашем случае число значений x (с учетом группировки) составило k=22, один параметр был оценен по выборке, т.е. d=1, поэтому если H0 верна, то тестовая статистика подчиняется хиквадрат распределению с k-d-1 = 22-1-1 = 20 степенями свободы.

Критическая точка уровня значимости 0.01 равна 39.968463129386:

alpha <- 0.01; chi2cr <- qchisq(1-alpha/2, 20).



Поскольку расчетное значение тестовой статистики (82.4507694346386) превышает критическое (39.968463129386), нулевую гипотезу о соответствии логнормальному закону распределения отвергаем.



5.Построим таблицу сопряженности по переменным «sex» и «city» (рис. 8) и проверим гипотезу о том, что выборки являются статистически однородными. Воспользуемся функцией table:

t <- table(var3$sex,var3$city).

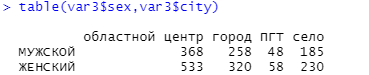


Рисунок 8 – Таблица сопряженности по переменным «sex» и «city»

Для проверки в данном примере воспользуемся встроенной функцией chisq.test, входящей в R: chisq.test(t).

Напомним, что в тесте на однородность выборок по таблице сопряженности рассматривается нулевая гипотеза о том, что выборки однородны. Альтернативная гипотеза: в выборках есть неоднородность. Результаты проведения теста представлены на рис. 9.

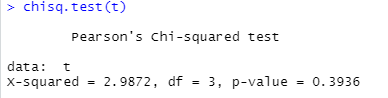


Рисунок 9 – Результаты теста хи-квадрат на однородность выборок

Поскольку p-value в нашем случае превышает уровень значимости 5%, можно сделать вывод о том, что нулевую гипотезу отвергать не следует, выборки однородны: доли мужского и женского населения в областном центре, городе, ПГТ и селе значимо не различаются.

Вывод: в результате проведения теста Колмогорова-Смирнова Н0 отвергается, соответствия логнормальному распределению нет. При получении значения тестовой статистики Пирсона и критической точки уровня значимости 0.01 сделали вывод о том, что нулевая гипотеза о соответствии логнормальному закону распределения отвергается. В результате проведения теста хи-квадрат на однородность выборок сделали вывод о том, что нулевую гипотезу отвергать не следует, выборки однородны.