МатСтат Дз

Состав группы: Сысоев Никита, Каменских Ирина, Сухов Александр

Задание 2

Среднее значение дохода и дисперсия

Среднее значение — 635.888, дисперсия — 104660.1

Асимметрия и Эксцесс

Асимметрия — 1.569469, эксцесс — 6.215781

На графике мы видим длинный хвост распределения справа, что говорит нам о том, что распределение скошено вправо. Тогда из теории мы знаем, что асимметрия должна быть положительна. Получив асимметрию из данных, мы в этом убеждаемся. Значение асимметрии > 0.5 => асимметрия существенная.
Из теории мы знаем, что положительный эксцесс говорит о том, что эмпирическое распределение является более высоким («островершинным») – относительно «эталонного» нормального распределения с параметрами $a = \bar{x}$ и $\sigma = s$. И чем больше эксцесс по модулю, тем «аномальнее» высота в ту или иную сторону. В нашем случае эксцесс положительный.

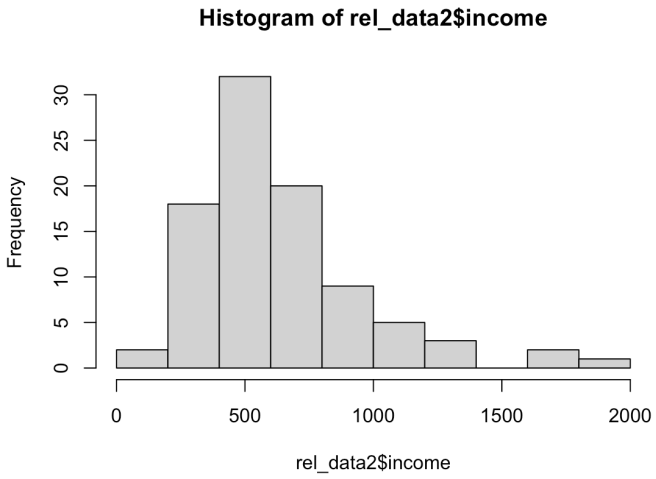

График 1. Гистограмма дохода

Сгенерировав выборку из нормального распределения с данными параметрами, получаем:

```
g <- rnorm(n=92, mean=mean_income_2,
sd=sqrt(var_income_2))
g
hist(g)
skewness(g) # 0.09766422
kurtosis(g) # 2.372802
```

Мы можем считать это распределение по выборке примерно эталонным. Действительно, сравнивая 2 графика, убеждаемся в том, что правый хвост распределения исследуемой выборки достаточно длинный, и исследуемая выборка более вытянутая вверх, чем представленная на графике 2.

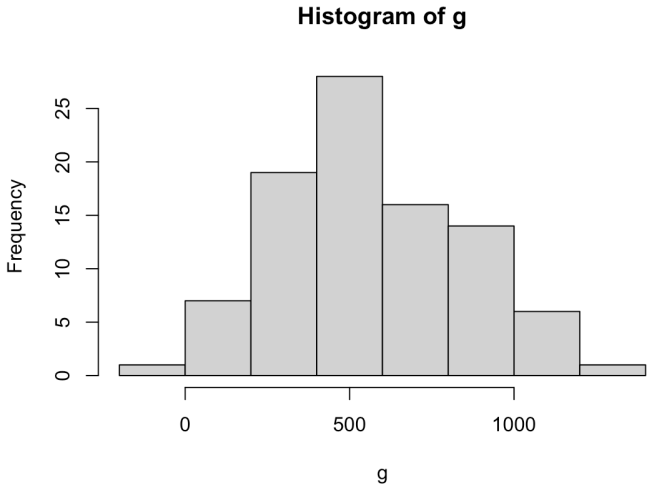

График 2. Гистограмма нормального распределения

Тест на нормальность дохода

Проводя тест на нормальность Шапиро и тест Колмогорова-Смирнова, получаем довольно разные p-value (3.778e-07 и 0.02239), но они < 0.05 => отвергаем нулевую гипотезу о нормальности дохода на уровне значимости 5% и выше.

Корреляция и значимость корреляции дохода и посевной площади

Проводя тесты на корреляцию, получаем, что во всех случаях (Пирсон, Кендалл, Спирмен) получились очень маленькие p-value (2.2e-16, 2.674e-14 и 2.2e-16) => отвергаем нулевую гипотезу => корреляция значимо отделена от 0. Оценки корреляции — 0.7408701, 0.5402764 и 0.7356443 => достаточно высокая корреляция

Примечание:

Выдержка из документации: If method is «Kendall» or «spearman», Kendall's τ or Spearman's ρ statistic is used to estimate a rank-based measure of association. These tests may be used if the data do not necessarily come from a bivariate normal distribution.

Поэтому результат полученный через Пирсона мы провели, но держим в уме, что его результаты могут быть неверными (т.к. мы отвергли гипотезу о нормальности данных выше).

Гипотеза о равенстве дисперсий дохода и расхода

Проводим var.test (он требует нормальности данных, но т.к. ничего другого на семинарах мы не прошли, то можно использовать его).
p-value > 0.05 (0.5713) => НЕ отвергаем нулевую гипотезу => предполагаем, что истинные дисперсии равны.

Доверительный интервал для средней стоимости скота

Т.к. в файле с Новгородской обл. нет данных по стоимости скота, мы решили, что есть смысл брать в качестве этого параметра колонку «Общая стоимость (без земли)», предполагая, что стоимость складывается из земли и скота.
Посевная площадь — сумма колонок «посевная площадь (надельная)», «посевная площадь (купчая)», «посевная площадь (арендованная, надельная)» и «посевная площадь (арендованная, купчая)»

Применяя функцию *eexp*, получаем ДИ для параметра λ показательного распределения. Т.к. мы ищем ДИ для мат. ожидания, к-рое равно обратной величине, перейдем к ДИ обратной величины.
Получаем: (858.791, 1293.749)

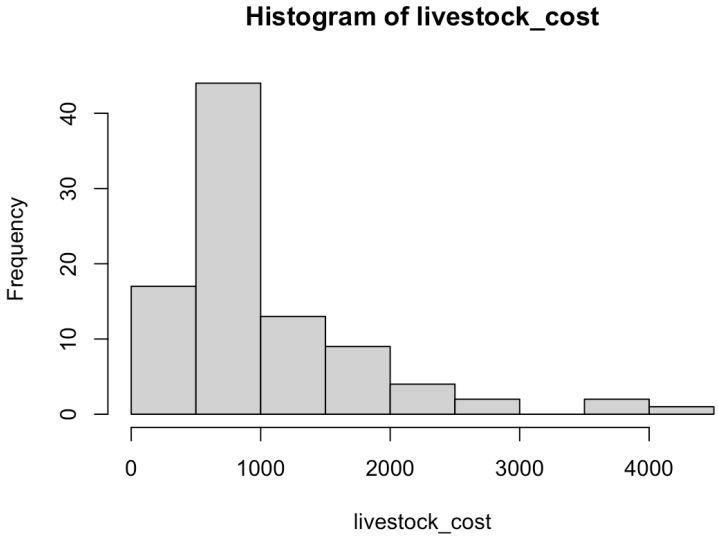

График 3. Гистограмма стоимости скота

Полученный результат не идет в разрез с наблюдаемой картиной.

$\overline{x}$ = 1042.944

Тот же результат получается при построении ДИ вручную (см. скрипт):

$$\xi \sim exp(\lambda)$$
$$2\lambda n\overline{x} \sim \chi^2_{2n}$$
$$\chi^2_{1-\frac{\alpha}{2},2n} < 2\lambda n\overline{x} < \chi^2_{\frac{\alpha}{2},2n}$$
$$\frac{2n\overline{x}}{\chi^2_{\frac{\alpha}{2},2n}} < \frac{1}{\lambda} < \frac{2n\overline{x}}{\chi^2_{1-\frac{\alpha}{2},2n}}$$

Задание 1

Среднее значение дохода и дисперсия

Среднее значение — 1912.269, дисперсия — 1171135

Асимметрия и Эксцесс

Асимметрия — 1.422833, эксцесс — 3.250623

На графике мы видим длинный хвост распределения справа, что говорит нам о том, что распределение скошено вправо. Тогда из теории мы знаем, что асимметрия должна быть положительна. Получив асимметрию из данных, мы в этом убеждаемся. Значение асимметрии > 0.5 => асимметрия существенная.
Из теории мы знаем, что положительный эксцесс говорит о том, что эмпирическое распределение является более высоким («островершинным») – относительно «эталонного» нормального распределения с параметрами $a = \overline{x}$ и $\sigma = s$. И чем больше эксцесс по модулю, тем «аномальнее» высота в ту или иную сторону. В нашем случае эксцесс положительный.

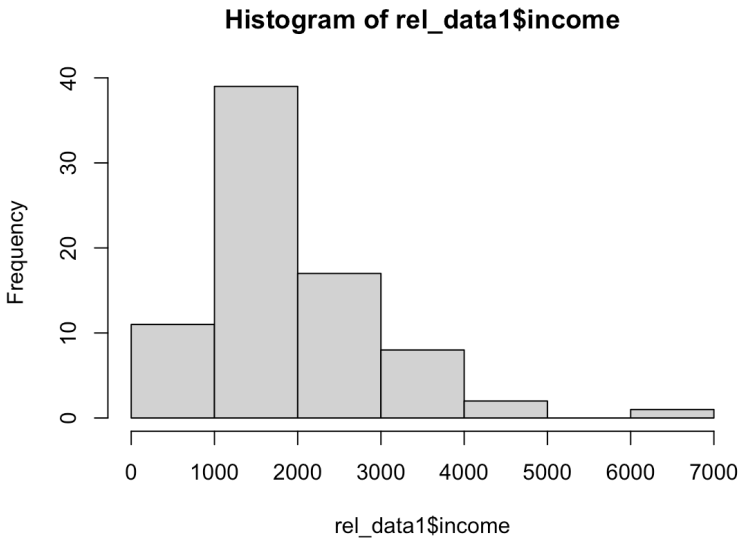

График 4. Гистограмма дохода

Сгенерировав выборку из нормального распределения с данными параметрами, получаем:

```
g <- rnorm(n=78, mean=mean_income_1, sd=sqrt(var_income_1))
g
hist(g)
skewness(g) # 0.1020892
kurtosis(g) # 0.1760094
```

Мы можем считать это распределение по выборке примерно эталонным. Действительно, сравнивая 2 графика, убеждаемся в том, что правый хвост распределения исследуемой выборки достаточно длинный, и исследуемая выборка более вытянутая вверх, чем представленная на графике 2.

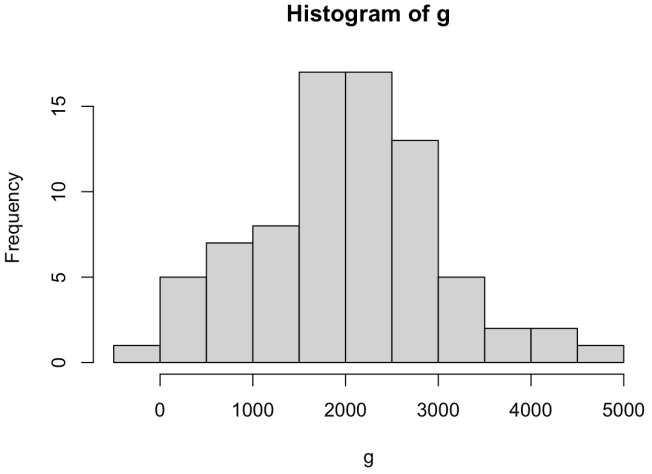

График 5. Гистограмма нормального распр.

Тест на нормальность дохода

Проводя тест на нормальность Шапиро и тест Колмогорова-Смирнова, получаем маленькое p-value (1.955e-05 и 0.08093), одно больше 0.05, другое меньше, но тест Шапиро точнее, поэтом будем ориентироваться на него => отвергаем нулевую гипотезу о нормальности дохода на уровне 10% и выше уверенно и с чуть меньшей уверенностью на уровне значимости 5%.

Корреляция и значимость корреляции дохода и посевной площади

Примечание: В качестве посевной площади рассматривается колонка «земли пахотной всего».

Проводя тесты на корреляцию, получаем, что во всех случаях (Пирсон, Кендалл, Спирмен) получились очень маленькие p-value (1.188e-12, 3.919e-10 и 2.51e-10) => отвергаем нулевую гипотезу => корреляция значимо отделена от 0.
Оценки корреляции — 0.6982993, 0.484223, 0.641372 => достаточно высокая корреляция.

Примечание:

Выдержка из документации: If method is «Kendall» or «spearman», Kendall's τ or Spearman's ρ statistic is used to estimate a rank-based measure of association. These tests may be used if the data do not necessarily come from a bivariate normal distribution.
Поэтому результат полученный через Пирсона мы провели, но держим в уме, что его результаты могут быть неверными (т.к. мы отвергли гипотезу о нормальности данных выше).

Гипотеза о равенстве дисперсий дохода и расхода

Проводим var.test (хотя он требовательный к нормальности данных).
p-value > 0.05 (0.6986) => НЕ отвергаем нулевую гипотезу => предполагаем, что истинные дисперсии равны.

Доверительный интервал (95%) для средней стоимости скота

Применяя функцию *eexp*, получаем ДИ для параметра λ показательного распределения. Т.к. мы ищем ДИ для мат. ожидания, к-рое равно обратной величине, перейдем к ДИ обратной величины.
Получаем: (782.5853, 1221.5159)