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.

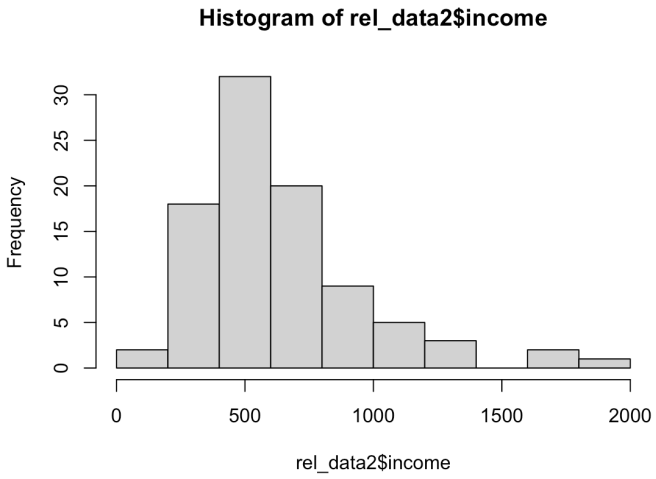


График 1. Гистограмма дохода

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.

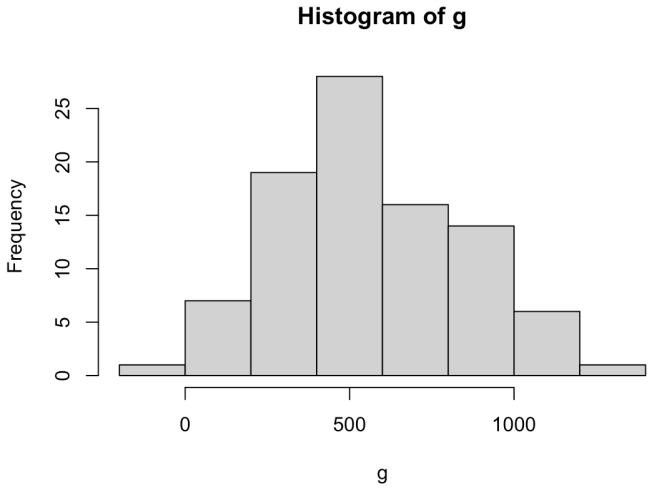


График 2. Гистограмма нормального распределения

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)

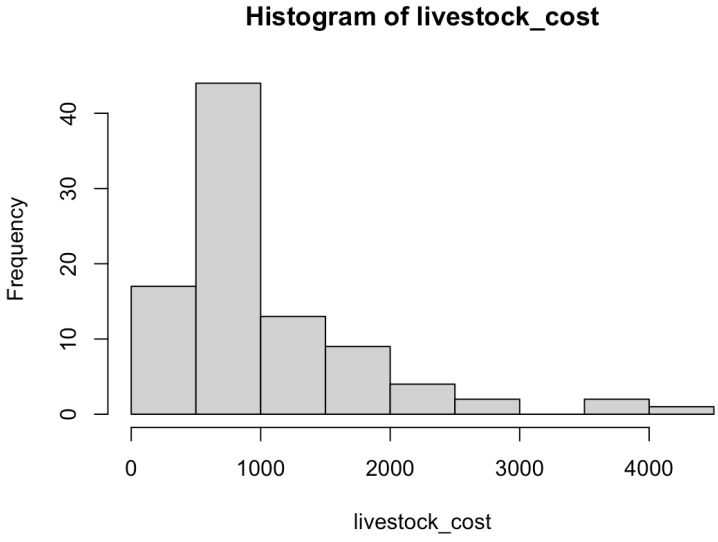


График 3. Гистограмма стоимости скота

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.

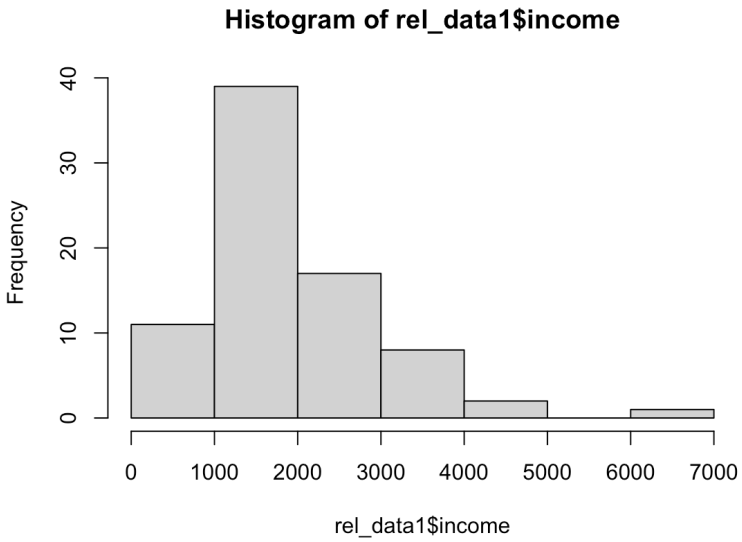


График 4. Гистограмма дохода

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.

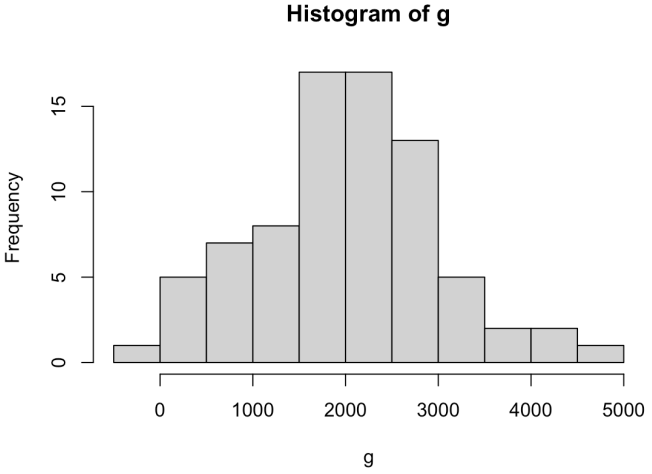


График 5. Гистограмма нормального распр.

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.

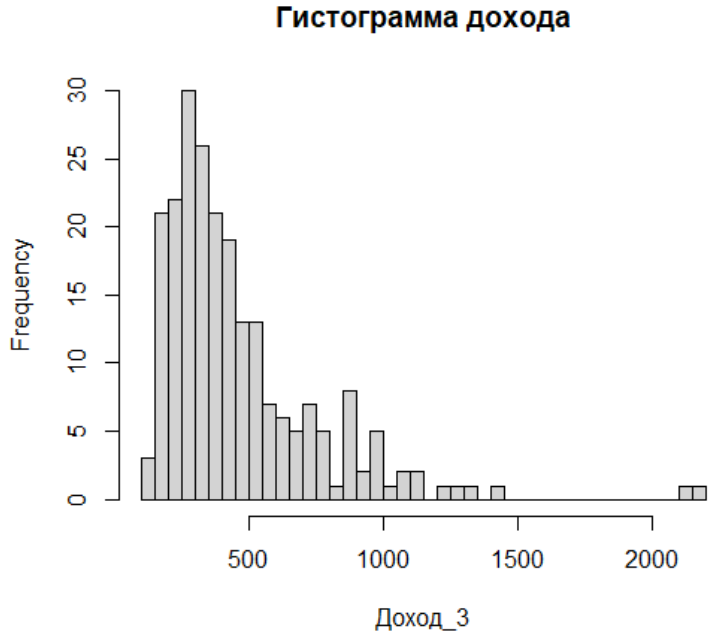
Доверительный интервал (95%) для средней стоимости скота

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)

Задание 3

1. Рассчитываем среднее значение и дисперсию дохода с помощью функций mean() и var().

   Среднее значение дохода = 464,242; дисперсия дохода = 90615,324.

2. 

3. Рассчитываем асимметрию и эксцесс с помощью функций skewness() и kurtosis(). Значение асимметрии = 2,289; значение эксцесса = 11,09997.

   Асимметрия больше нуля, следовательно, распределение смещено вправо, имеет длинный хвост справа.

   Эксцесс положителен, что означает, что выбросы в данных интенсивнее, чем для нормального распределения. Эмпирическое распределение является более 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.

4. Тест на нормальность дохода: используем shapiro.test() и ks.test() (Тест Колмогорова-Смирнова проводим, сравнивая нашу выборку с выборкой из нормальной генеральной совокупности с параметрами $a=\bar{x}$ и $\sigma=S$).

   В первом случае получаем p-value = 2,562e-16, во втором: p-value = 2.1e-05. p-value достаточно маленькие (p-value < 0,05), следовательно, отвергаем нулевую гипотезу о том, что значения дохода распределены нормально, на уровне значимости 5% и выше.

5. Корреляция дохода и посевной площади, её значимость: проводим тесты на корреляцию методами spearman, kendall (метод pearson не используем, так как в предыдущем пункте опровергли гипотезу о нормальности распределения дохода). Получаем p-value < 2.2e-16; p-value < 2.2e-16 соответственно. Значения p-value достаточно маленькие, следовательно, отвергаем нулевую гипотезу, и корреляция значимо отделена от 0.
   Оценки корреляции: 0,7516853, 0,5514624; имеем достаточно высокую корреляцию между доходом и посевной площадью.
6. Гипотеза о равенстве дисперсий дохода и расхода: проводим var.test(). (предположим, что данные распределены нормально). p-value = 0,001524 (p-value < 0,05), следовательно, отвергаем нулевую гипотезу и предполагаем, что истинные дисперсии не равны, на уровне значимости 5% и выше.
7. Доверительный интервал для средней стоимости скота:

   Применяем функцию eexp(), получаем доверительный интервал для параметра показательного распределения. Чтобы найти доверительный интервал для мат. ожидания, которое равно 1/λ, поделим 1 на границы найденного доверительного интервала и поменяем эти значения местами (1/UCL<1/LCL). Получаем: (354; 461).

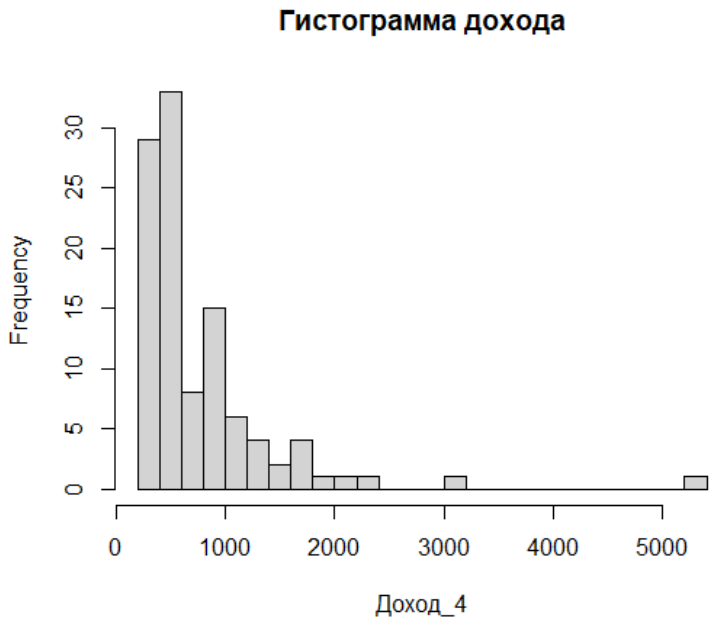
   Результат совпадает с доверительным интервалом уровня значимости 5%, полученным следующим преобразованием (предполагаем, что стоимость скота имеет показательное распределение):

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

Задание 4

1. Рассчитываем среднее значение и дисперсию дохода с помощью функций mean() и var().

   Среднее значение дохода = 758.927; дисперсия дохода = 430690.803.

2. 

3. Рассчитываем асимметрию и эксцесс с помощью функций skewness() и kurtosis(). Значение асимметрии = 3,685; значение эксцесса = 22,578.

   Асимметрия больше нуля, следовательно, распределение смещено вправо, имеет длинный хвост справа.

   Эксцесс положителен, что означает, что выбросы в данных интенсивнее, чем для нормального распределения. Эмпирическое распределение является более 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.

4. Тест на нормальность дохода: используем shapiro.test() и ks.test() (Тест Колмогорова-Смирнова проводим, сравнивая нашу выборку с выборкой из нормальной генеральной совокупности с параметрами $a=\bar{x}$ и $\sigma=S$).

   В первом случае получаем p-value = 2.323e-14, во втором: p-value = 0.0004466. p-value достаточно маленькие (p-value < 0,05), следовательно, отвергаем нулевую гипотезу о том, что значения дохода распределены нормально, на уровне значимости 5% и выше.

5. Корреляция дохода и посевной площади, её значимость: проводим тесты на корреляцию методами spearman, kendall (метод pearson не используем, так как в предыдущем пункте опровергли гипотезу о нормальности распределения дохода). Получаем p-value = 4.886e-12; p-value = 9.74e-11 соответственно. Значения p-value достаточно маленькие, следовательно, отвергаем нулевую гипотезу, и корреляция значимо отделена от 0.
   Оценки корреляции: 0,6077856; 0,4290236.

   Примечание: посевной площадью считаем сумму площадей под озимую пшеницу и под яровую пшеницу, так как посевной площадью называется часть пахотных земель, занятых посевами сельскохозяйственных культур. Нам известна площадь общая площадь пашни, однако, имеем недостаточно данных, чтобы определить, засевается вся эта общая площадь или нет.

6. Гипотеза о равенстве дисперсий дохода и расхода: проводим var.test() (предположим, что данные распределены нормально). p-value < 2.2e-16 (p-value<0,05), следовательно, отвергаем нулевую гипотезу и предполагаем, что истинные дисперсии не равны на уровне значимости 5% и выше.
7. Доверительный интервал для средней стоимости скота:

   Применяем функцию eexp(), получаем доверительный интервал для параметра показательного распределения. Чтобы найти доверительный интервал для мат. ожидания, которое равно 1/λ, поделим 1 на границы найденного доверительного интервала и поменяем эти значения местами (1/UCL<1/LCL). Получаем: (307; 451).

   Результат совпадает с доверительным интервалом уровня значимости 5%, полученным следующим преобразованием (предполагаем, что стоимость скота имеет показательное распределение):

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

Задание 5

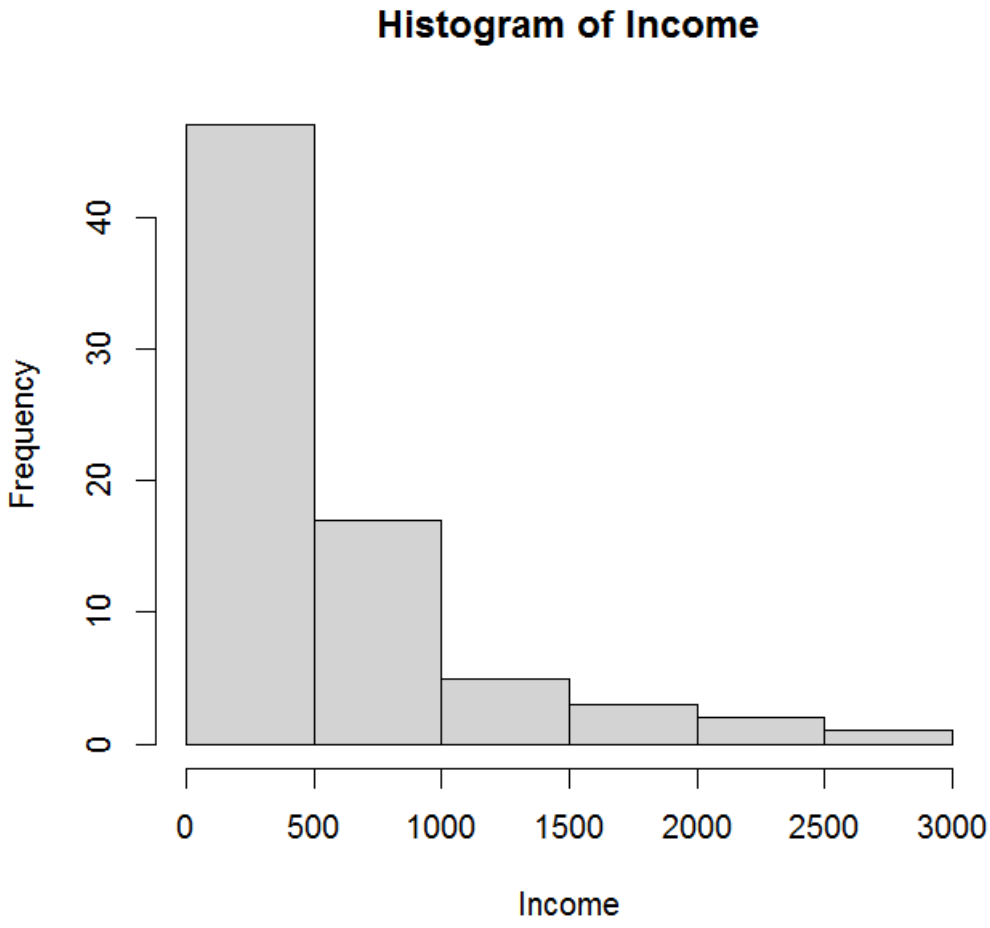
№1

Для начала удалим пустую строку из файла эксель. Будем считать, что общий доход является суммой валового дохода от земледелия и полеводства и дохода от скотоводства и птицы.

В столбце валового дохода заменим пропуски на нули, ни среднее ни дисперсия не изменятся.

Среднее значение дохода – 586.59, дисперсия – 313737.4.

№2



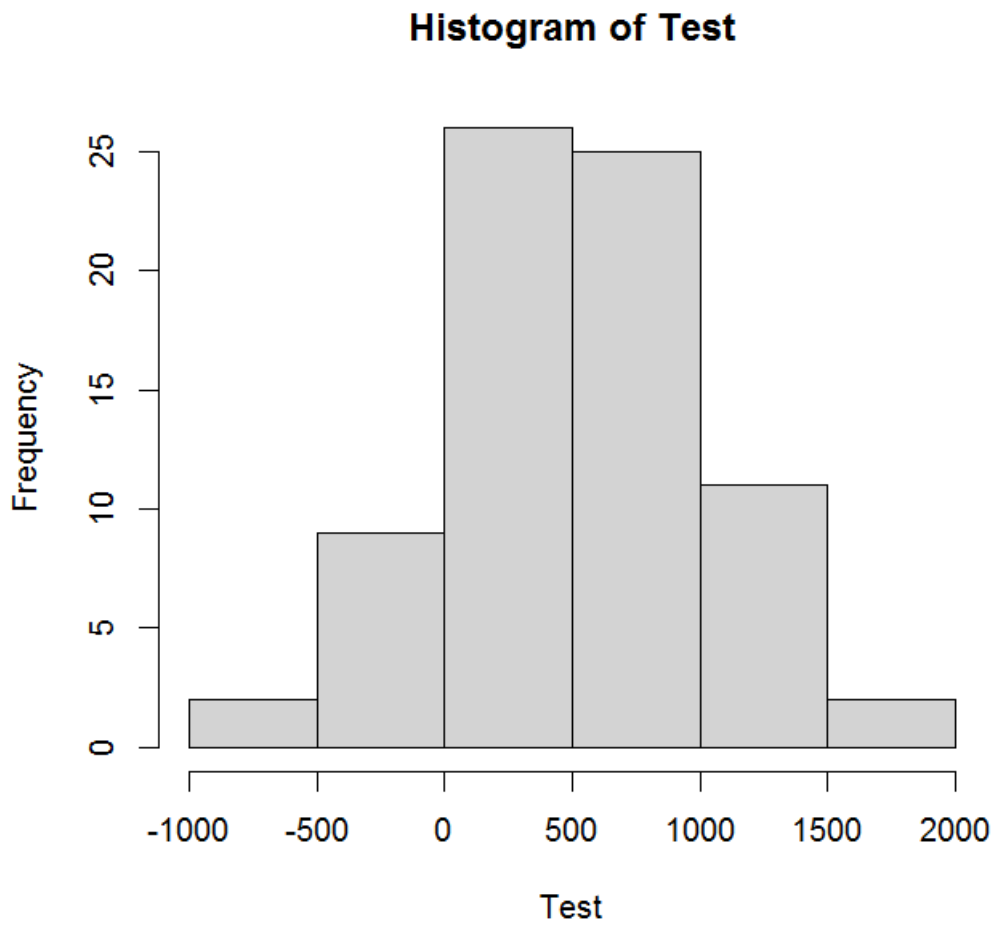
№3

Асимметрия – 2.123, эксцесс – 7.85

Асимметрия характеризует меру скошенности, что заметно на гистограмме – распределение скошено вправо. Поэтому асимметрия больше нуля. В случае, если наблюдается скошенность влево, асимметрия будет отрицательной. Чем больше по модулю асимметрия, тем более значительна скошенность в соответствующую сторону.

Эксцесс характеризует высоту или остроконечность распределения. Чем больше эксцесс, тем выше верхняя точка распределения и наоборот.

Для сравнения сгенерируем случайную выборку с тем же числом элементов и с нашими найденными средним значением и дисперсией.



Асимметрия – 0.123, эксцесс – 2.68

Сравнив частоту наиболее часто встречающихся величин заметим, что у исходной выборки эксцесс действительно больше, и наблюдается правосторонняя асимметрия, в то время как у сгенерированной выборки распределение практически симметрично.

№4

Проводим тесты Шапиро-Уилка и Колмогорова-Смирнова, получая следующие p-value: 1.161e-09 и 0.00075 соответственно. Следовательно, поскольку оба значения меньше 0.05, мы не можем принять гипотезу о нормальности распределения дохода.

№5

За посевную площадь будем считать сумму площадей долевых и арендованных полей.

Распределение дохода не нормально. Критерий Пирсона основывается на нормальности данных, поэтому им мы пользоваться не будем.

Согласно методу Спирмена, корреляция равна 0.812, p-value = 2.2e-16

Согласно методу Кендалла, корреляция равна 0.646, p-value = 7.089e-16

Нулевая гипотеза – равенство корреляции нулю, однако оба значения p-value крайне малы, поэтому мы отвергаем нулевую гипотезу о равенстве корреляции дохода и посевной площади нулю.

№6

Будем считать, что суммарные расходы складываются из расходов на личные и хозяйственные потребности.

Предположим, что исследуемые данные распределены нормально (необходимо для var.test). Получим p-value = 0.001194 < 0.05, значит нужно отвергнуть нулевую гипотезу о равенстве дисперсий. Дисперсии доходов и расходов не равны.

№7

Удалим из вектора стоимости скота ячейку, где нет значения (сам вектор является соответствующим столбцом из экселя).

Воспользуемся следующим преобразованием, предполагая показательность распределения:

$$\xi \sim exp(\lambda)$$

$$2\lambda n\bar{x} \sim \chi^2_{2n}$$

$$\chi^2_{1-\frac{\alpha}{2},2n} < 2\lambda n\bar{x} < \chi^2_{\frac{\alpha}{2},2n}$$

$$\frac{2n\bar{x}}{\chi^2_{\frac{\alpha}{2},2n}} < \frac{1}{\lambda} < \frac{2n\bar{x}}{\chi^2_{1-\frac{\alpha}{2},2n}}$$

Нам необходимо найти лишь значения хи-квадрат в критических точках и посчитать среднее по стоимости скота – 382.
Подставив значения получаем доверительный интервал для 1/λ, т.е. для мат. ожидания показательного распределения. Асимптотический доверительный интервал равен (307.98, 486.49).