Отчет учащегося 3 курса

Факультете компьютерных технологий и

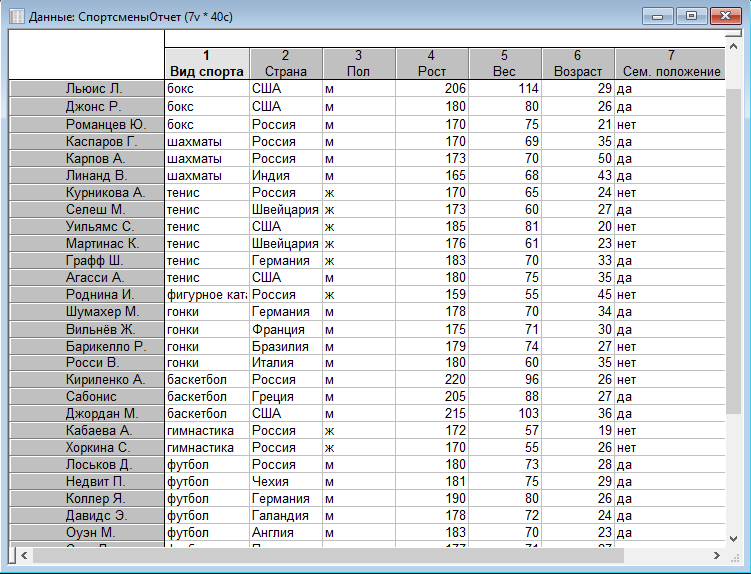
Прикладной Математики

Студента 37 группы

Павлова Дмитрия Павловича

Краснодар 2023

В данной работе была использована следующих таблиц сходных данных:



В которой имеются следующие переменные:

Вид спорта – категориальная переменная , указывающая каким видом спорта занимается спортсмен.

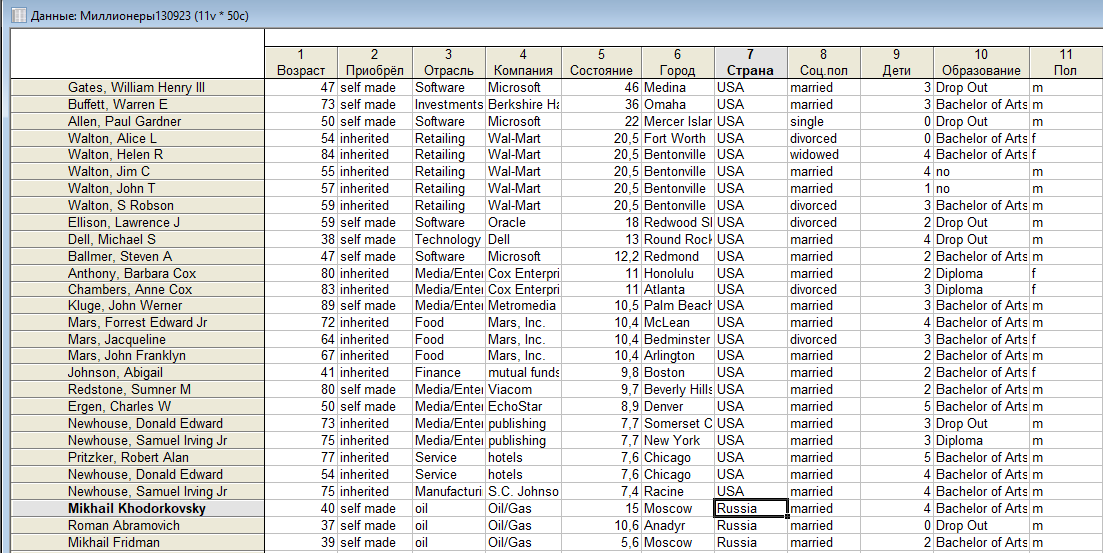
Страна – категориальная переменная , указывающая к какой стране относится спортсмен .

Пол- категориальная переменная , указывающая пол спортсмена

Рост – непрерывная , числовая переменная , указывающая рост спортсмена .

Вес – числовая переменная , указывающая вес спортсмена

Семейное положение – категориальная переменная , характеризующая семейное положение .



В данной таблице исходных данных имеются следующие переменные:

Возраст – числовая переменная , указывающая возраст миллионера на момент измерения

Приобрел – категориальная переменная , указывающая как была получена компания (создал сам либо уначледовал).

Отрасль – категориальная переменная , указывающая отрасль деятельности .

Компания – категориальная переменная , указывающая компанию с которой связан миллионер .

Состояния – числовая переменная , указывающая состояние миллионера в миллионах долларов .

Город – текстовая переменная , указывающая город происхождения миллионера .

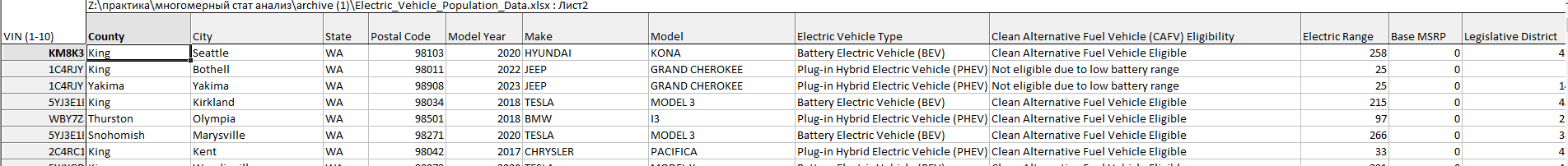
Страна – категориальная переменная , указывающая из какой страны миллионер .

Социальное положение – категориальная переменная , указывающая семейное положение .

Дети – числовая переменная , указывающая количество детей

Образование – категориальная переменная , указывающая образование миллионера

Пол – категориальная переменная , указывающая пол миллионера



Изображение выглядит как текст, Шрифт, число, линия

Автоматически созданное описание

В данной таблице исходных данных имеются следующие переменные:

Country – категориальная переменная , указывающая страну автомобиля

City – текстовая переменная , указывающая город автомобиля

State – категориальная переменная , указывающая состояние автомобиля

postal code – числовая переменная , указывающая почтовый индекс автомобиля

Model Year – числовая переменная , указывающая год выпуска автомобиля

Make – категориальная переменная , указывающая производителя автомобиля

Model – текстовая переменная , указывающая модель автомобиля

Electric Vehicle Type – категориальная переменная , указывающая тип электромобиля

Clean Alternative Fuel Vehicle (CAFV) Eligibility – категориальная переменная , указывающая право на использование экологически чистых транспортных средств на альтернативном топливе (CAFV)

Electric Range – числовая переменная , указывающая электрический диапазон

Base MSRP – числовая переменная , указывающая рекомендуемую розничную стоимость

Legislative District – rчисловая переменная указывающая законодательный округ

DOL Vehicle ID – числовая переменная , указывающая DOL идентификатор автомобиля

Vehicle Location – непрерывная переменная , указывающая местоположение автомобиля

Electric Utility – категориальная переменная , указывающая тип электроснабжения автомобиля

2020 Census Tract – числовая переменная , указывающая место в рейтинге в 2020 году

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

На данном слайде показана таблица корреляций между переменными таблицы спортсмены .

Чем ближе коэффициент корреляции к единице по модулю , тем сильнее корреляция между переменными .

Красным цветом в ячейках таблицы отмечены статистически значимые коэффициенты корреляции .

Из таблицы видно , что наибольшая взаимосвязь имеется между переменными рост и вес . Так как коэффициент корреляции у этих переменных больше нуля , увеличение или уменьшение значения одной переменной , приводит к увеличению (уменьшению ) значения другой . В случае если коэффициент корреляции меньше нуля, увеличение (уменьшение ) значения одной , приводит к уменьшению (увеличению) значения другой .

Наибольший по модулю , но отрицательный по знаку коэффициент корреляции имеют переменные Вес и Пол .

Если коэффициент корреляции по модулю близок к нулю , то взаимосвязь между переменными слабая , либо вообще отсутствует . Такая ситуация наблюдается между переменными Вид спорта и семейное положение .

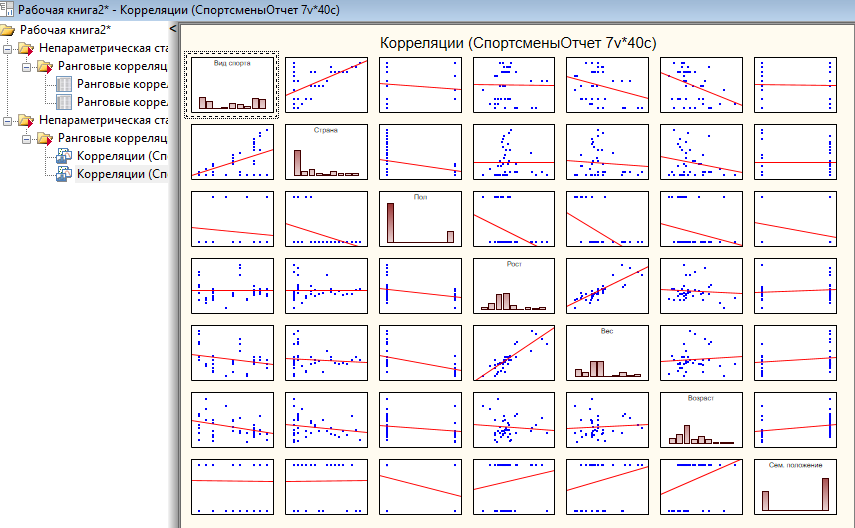
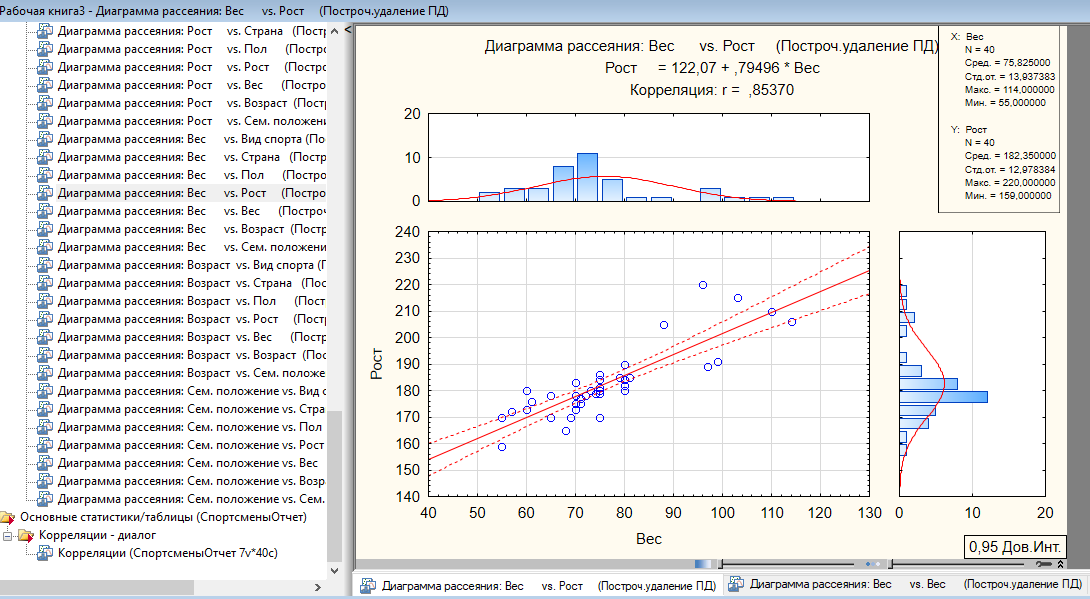


График корреляций между переменными

На картинках отображаются графики корреляций между переменными . Чем ближе синие точки к красной корреляционной линии , тем сильнее взаимосвязь . Если корреляционная линия убывает , то коэффициент корреляции меньше нуля , если возрастает , то больше .



На слайде показана диаграмма рассеивания для переменных Вес и Рост .

Верхний график отображает соотношение прогнозируемого (красная линия ) и фактического (Синие гистограммы ) значений для переменной Вес . Аналогичные диаграммы строятся для переменной Рост (Расположен под таблицей ) .

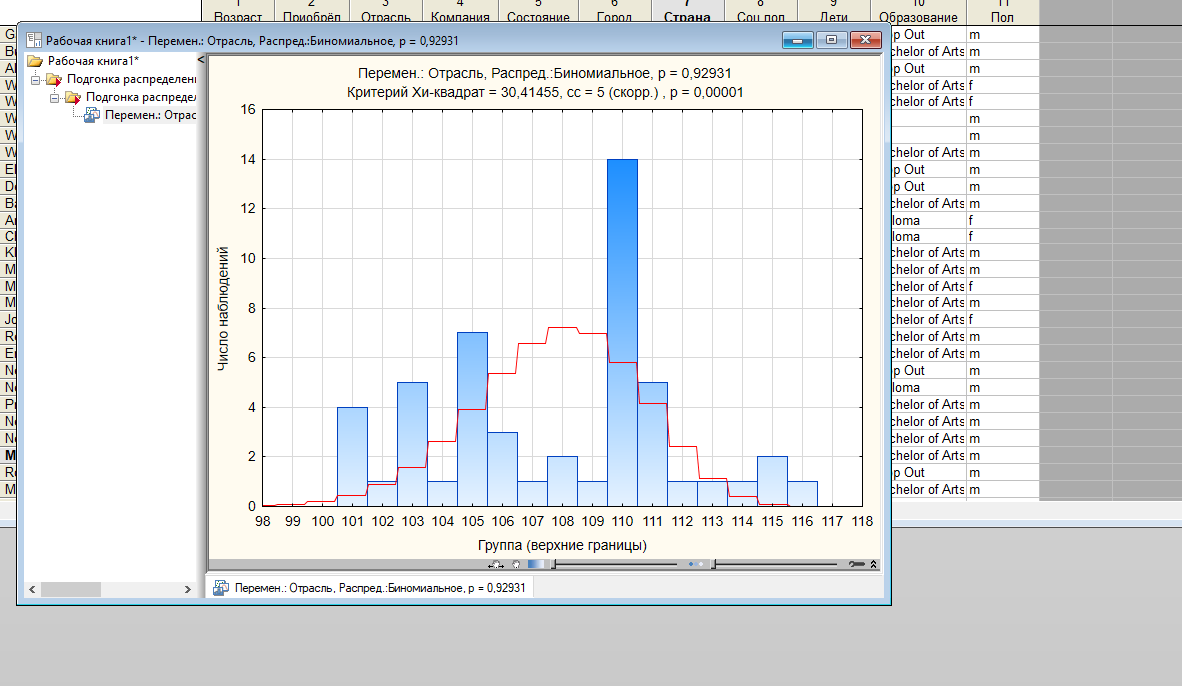
График расположенный между диаграммами , отображает коэффициент корреляции между переменными Рост и вес . Также , на графике присутствуют пунктирные линии , отображающие доверительный интервал , отклонение значений от корреляционной прямой в пределах которого является допустимым .

Над диаграммой веса , указывается коэффициент корреляции r , а над ним , формула взаимосвязи значений переменных рост и вес . В таблице приведены дополнительные сведения для данных переменных .

Так , это общее количество записей в таблице для каждой из переменных , среднее значение для каждой из них . Также , это величина стандартного отклонения , а также минимальное и максимальное значение .

## Подгонка распределения

Пакет statistica позволяет проверить соответствие исходных данных некоторому закону распределения .



Так , на приведенном слайде , показана диаграмма подгонки на соответствие биномиальному закону распределения для исходных данных таблицы миллионеры. Вверху указывается переменная , значения которой проверяются на соответствие выбранному закону распределения , название закона распределение ,значение критерия “Хи-Квадрат” указывающие , насколько переменная соответствует выбранному закону , чем меньше значение данного критерия ,тем выше вероятность соответствия переменной данному закону .

Также присутствует параметр числа степеней свободы cc и уровня значимости p.

Красной линией отмечены прогнозируемые значения , а синим фактические .

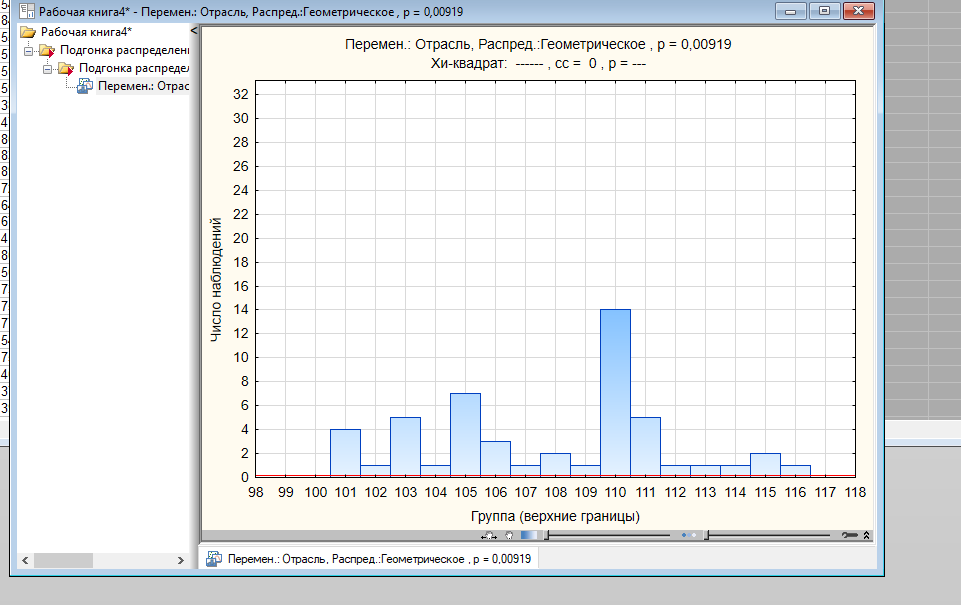
Из графика видно , что значения указанной переменной , не могут с достаточной точность описываться биномиальным законом распределения , так как присутствуют значительные отклонения в диапазоне от 106 до 111,а также от 115.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

На данном слайде показано соответствие значений переменной отрасль закону распределения Пуассона.

Из значения критерия “Хи Квадрат” можно сделать вывод , что данный закон распределения намного хуже описывает указанную переменную ( 67,86 против 30,41 у биномиального) . Диаграмма это также подтверждает . Так , из диаграммы видно, что все значения числа наблюдений , которые превышают значение 2 , не попадают в красную линию закона распределения , что создает значительные несоответствия прогнозируемых значений .



На данном слайде , показано соответствие значений переменной отрасль Геометрическому закону распределения , отсутствие критериев , говорит о невозможности описания данным распределением значений переменной .

Далее , будут приведены наблюдения подгонки распределения для переменной Состояние.

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

На слайде показано соответствие значений переменной Состояние нормальному закону распределения .

Из диаграммы можно видеть , что данному закону соответствуют группы с числом наблюдений от нуля до 10 .

Изображение выглядит как текст, программное обеспечение, Значок на компьютере, число

Автоматически созданное описание

В таблице показаны ожидаемые и фактические частоты для данного закона распределения .

Из таблицы видно , что максимальная разница между фактическим и предсказанным значением составляет 13,72 (для фактического 22 и прогнозируемого 8,27101) . Минимальная разница составляет 0,018 (для 0 и 0,018).

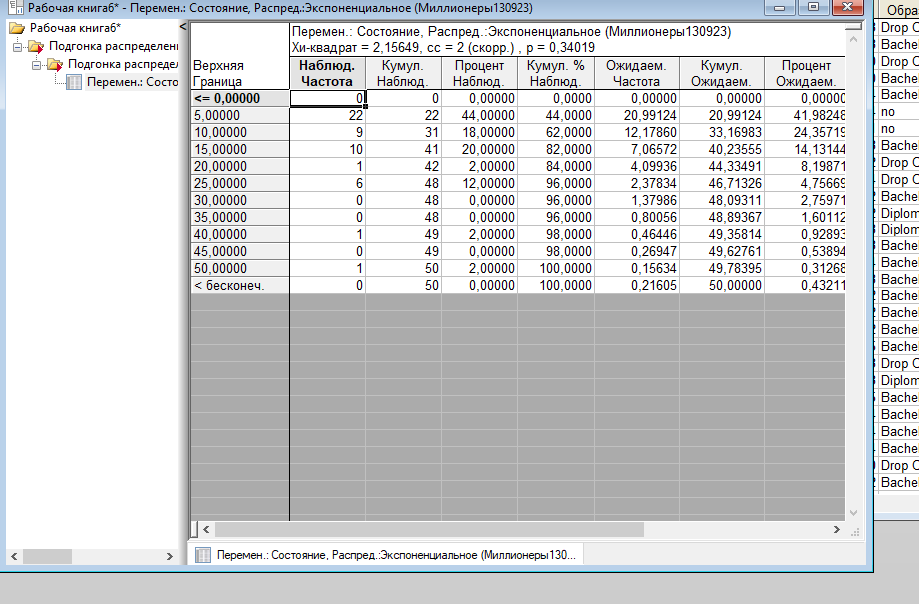
Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

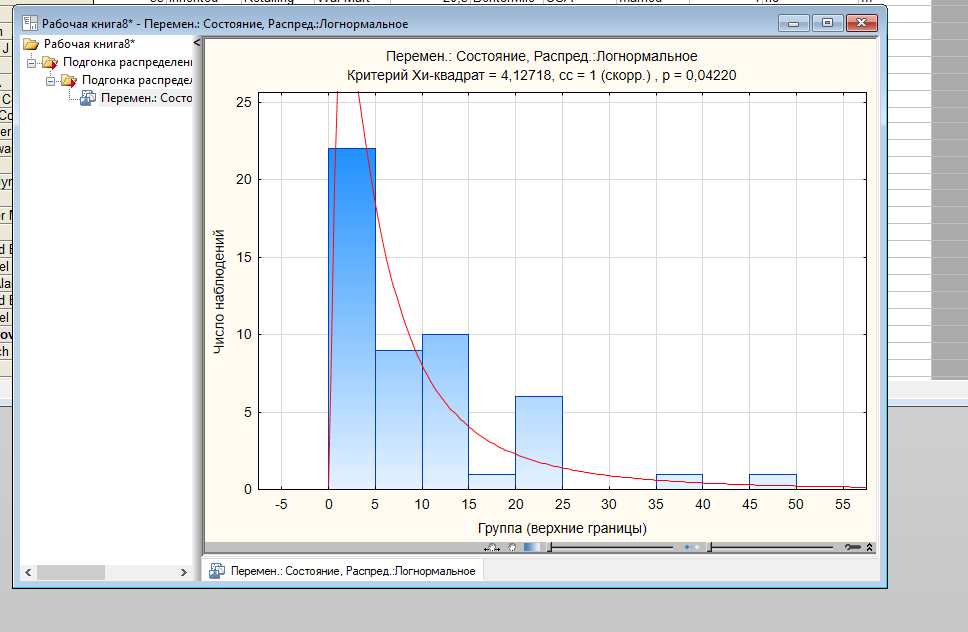
На слайде показано соответствие значений переменной Состояние экспоненциальному закону распределения . Из диаграммы видно , что данный закон наиболее хорошо описывает данную переменную .

Так , он учитывает большое число наблюдений в первой группе и дальнейшие снижения числа наблюдений .

В данном случае , часть диаграмм отсеченная красной линией сверху весьма мала , что указывает на низкий уровень ошибок при прогнозировании значений .



В данной таблице показаны числовые характеристики наблюдаемых и ожидаемых значений частот наблюдения , для данного закона распределения . Из таблицы видно , что максимальная разница между ожидаемой и наблюдаемой частотой составляет 3,63 ( для наблюдаемого 6 и ожидаемого 2,37) . Минимальная разница составляет 0,269 (наблюдаемая 0 , ожидаемая 0,269) .



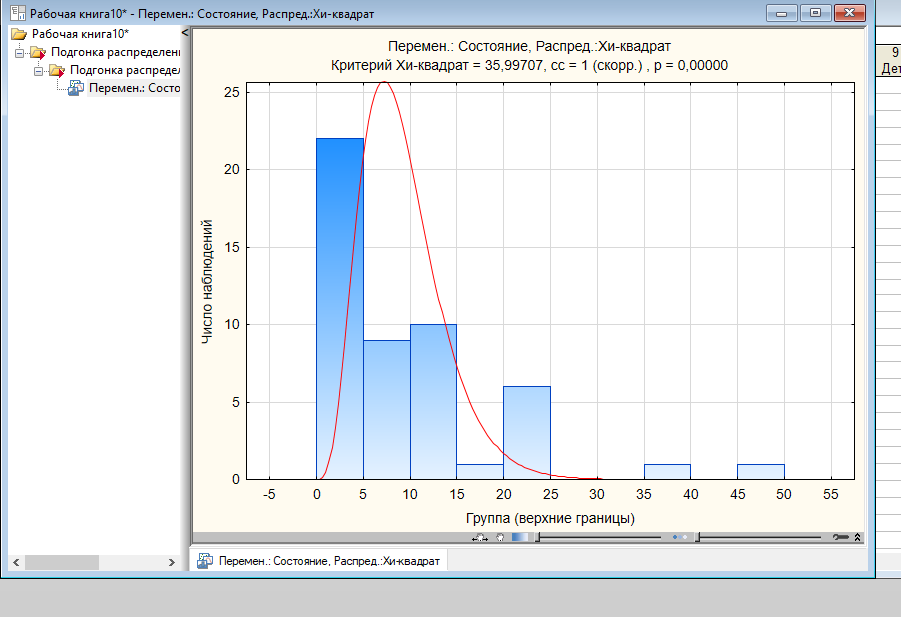
На данном слайде показано соответствие значений переменной Состояние логнормальному закону распределения .

Данный закон , также способен описывать данную переменную , однако , в отличии от экспоненциального , его прогнозируемые значения являются смещенными , что приводит к увеличению числа ошибок .

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

Из таблицы видно , максимальная разница между предсказанным и фактическим значением составляет 4,26152 (фактическое 10 , предсказанное 5,73848) . Минимальное значение составляет 0,365 (для 0 и 0,365 соответственно) , что выше минимального значения у экспоненциального распределения .



На данном слайде показано соответствие значений переменной состояние распределению “Хи квадрат”.

В отличии от 2 предыдущих , данной распределение описывает переменную намного хуже . Максимальное прогнозируемое значение смещено относительно фактического максимального значения .

Изображение выглядит как текст, программное обеспечение, число, Значок на компьютере

Автоматически созданное описание

Это также проявляется и в таблице . Так предсказываемые значения почти обращаются в ноль начиная со

строки 30 , хотя фактически нулями не являются ( что заметно и на графике ) .

Максимальная разница между предсказанным и фактическим составляет 14,30967 (для 22 и 7 ) , это объясняет смещение максимального прогнозируемого значения на графике . Минимальная разница составляет 0,003 (для 0 и 0,003 соответственно).

Итоговая таблица

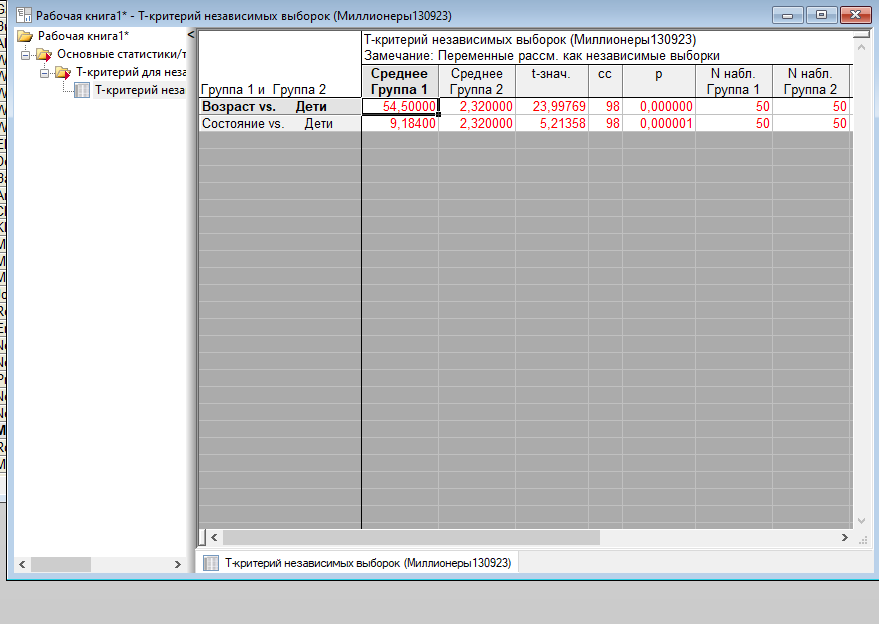
В таблице ниже приведены значения критерия “Хи квадрат ”, минимальной и максимальной разницы в прогнозируемом и предсказанном значение , а также название закона распределения для переменной Состояние .

|  |  |  |  |
| --- | --- | --- | --- |
| Название закона | Критерий “Хи квадрат” | Минимальная разница между предсказанным и фактическим значением | Максимальная разница между предсказанным и фактическим значением |
| Нормальное | 36,96 | 13,72 | 0,018 |
| Экспоненциальное | 2,15 | 3,63 | 0,269 |
| Логнормальное | 4,127 | 4,261 | 0,365 |
| Хи Квадрат | 35,997 | 14,30 | 0,003 |

Из таблицы видно, что экспоненциальное распределение является наиболее оптимальным для данной переменной.

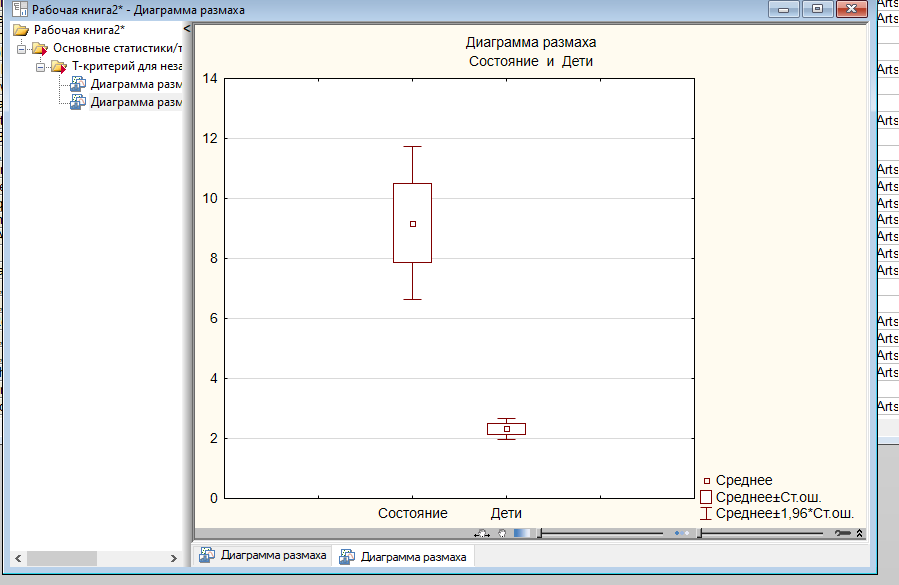
## Критерий Стьюдента равнения средних

Критерий Стьюдента используется для проверки гипотезы о равенстве средних двух выборок.



В таблице выше приведены результаты поверки Критерия Стьюдента (t критерия) для двух независимых переменных . В качестве группирующей переменной была выбрана переменная Дети .Разбиение на группы производилось программой автоматически. В таблице приведены значения среднего в обеих группах , также значение t критерия а также уровень значимости p , число наблюдений в обоих группах.

Чем более различны средние в группах, тем более сильная степень зависимости между независимой(группирующей ) и зависимыми переменными.

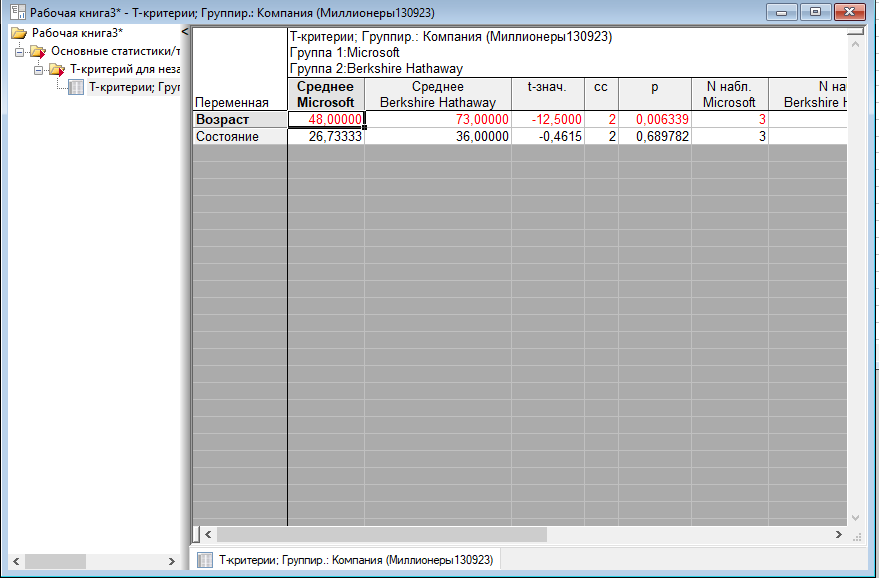


Слайд показывает диаграмму размаха для переменных состояние и дети.

На диаграмме изображено среднее значения для обоих переменных, а также стандартное отклонение и доверительный интервал отклонения , в пределах которого можно делать прогноз.

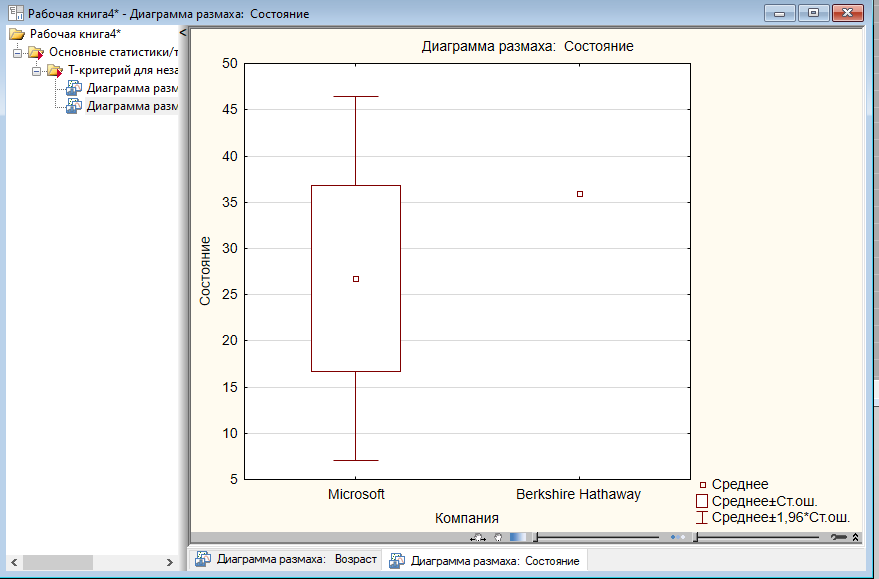
Помимо переменных , t критерий может быть вычислен для независимых групп .

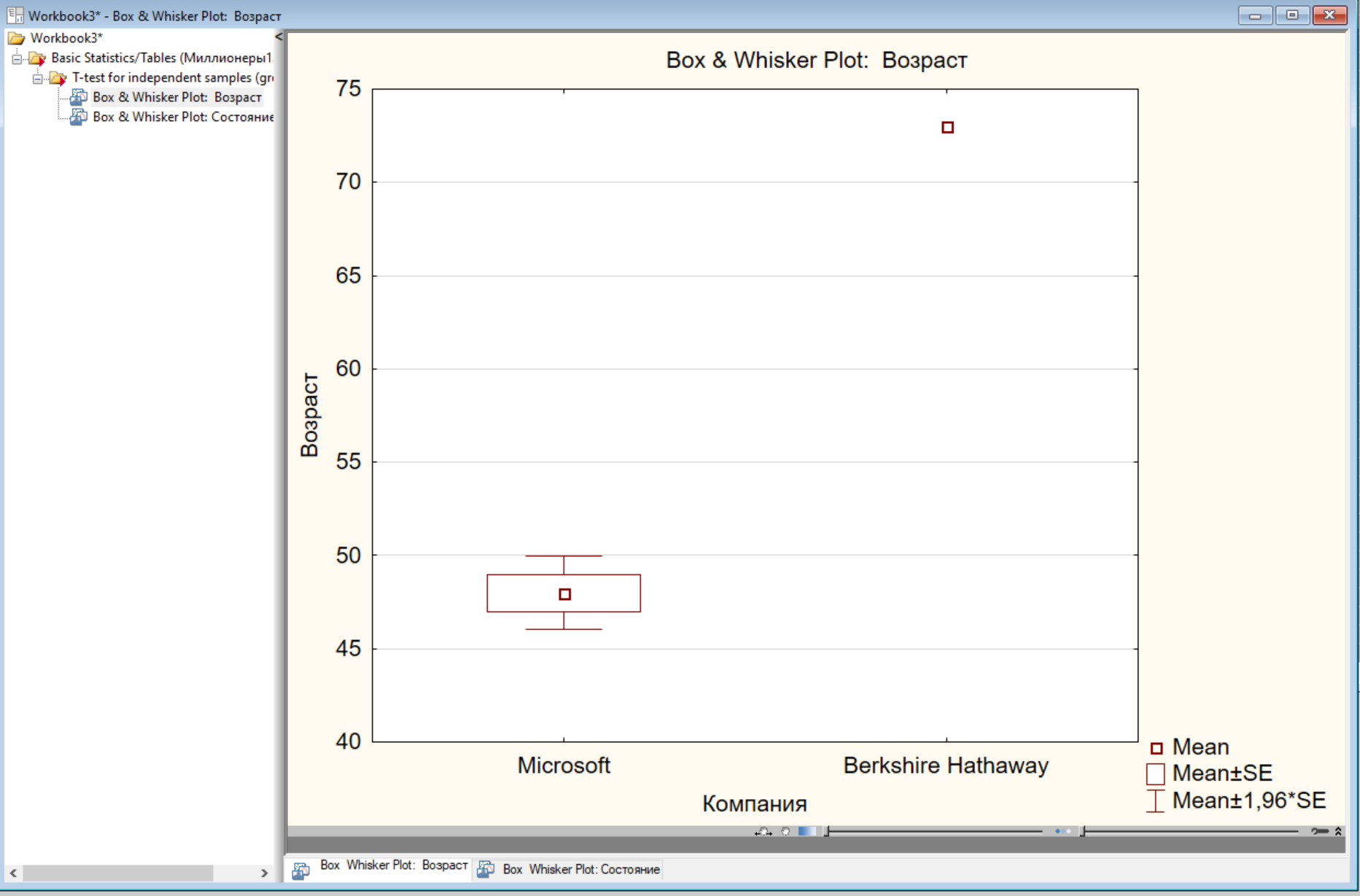
В этом случае ,в качестве группирующей переменной , должна быть выбрана категориальная переменная.



Здесь в качестве группирующей переменной , использовалась переменная компания и фактором влияющим на разбиение по группам являлись ее значения. В данном случае , статистически значимой является лишь связь переменных Возраст и Компания , так как для второй переменной уровень значимости p больше 0,05.

На следующих двух слайдах показаны диаграммы размаха для переменных “состояние и компания” и “возраст и компания”.





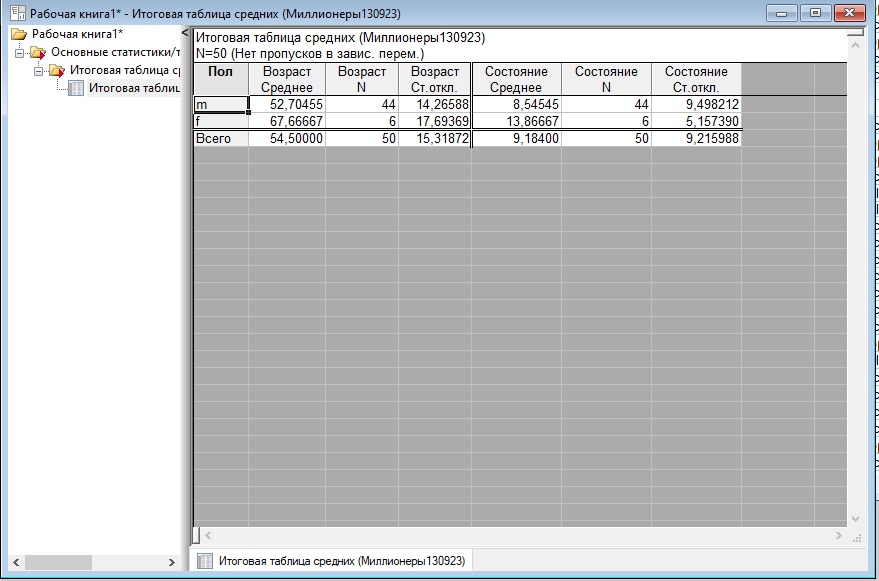
Из диаграммы видно ,что отклонение для состояния намного больше , чем отклонение для возраста .

## Однофакторный anova

Дисперсионный анализ более эффективен и информативен для малых выборок , чем t критерий.

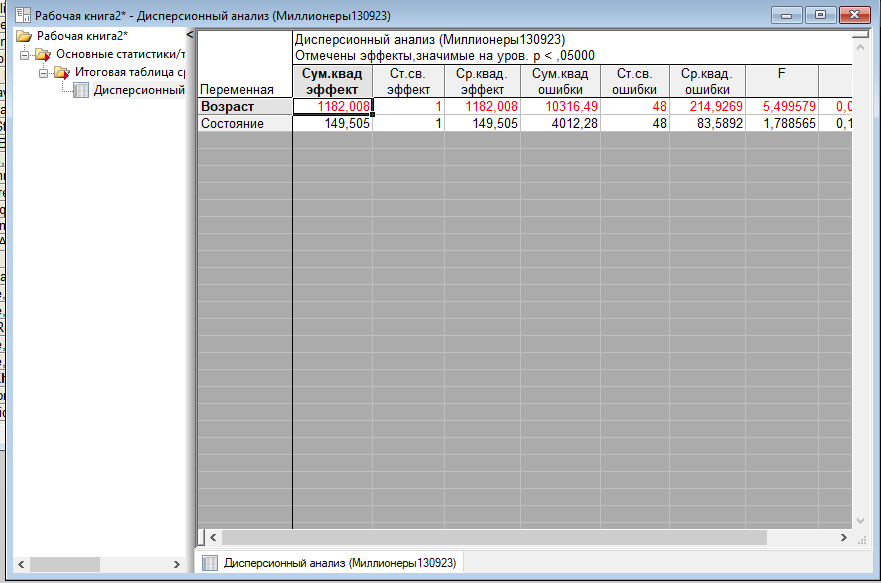
F критерий , применяемый в дисперсионном анализе , устойчив к отклонению от нормальности и однородности дисперсий.

Однофакторный дисперсионный анализ позволяет оценить эффект одной группирующей переменной (одного межгруппового фактора), на одну или более зависимых переменных.



На слайде показано влияние группирующей переменной пол , на зависимые переменные Возраст и Состояние.

Из таблицы видно, что число наблюдений в каждой группе N сильно отличается , что не позволяет делать достоверных выводов по результатам исследования.

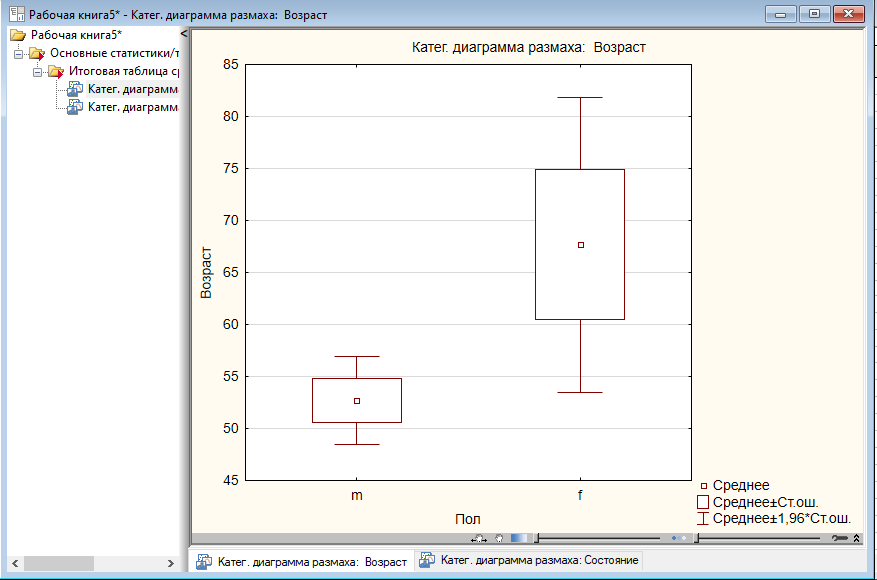


На слайде показан значения дисперсии ,ошибки и F критерия.

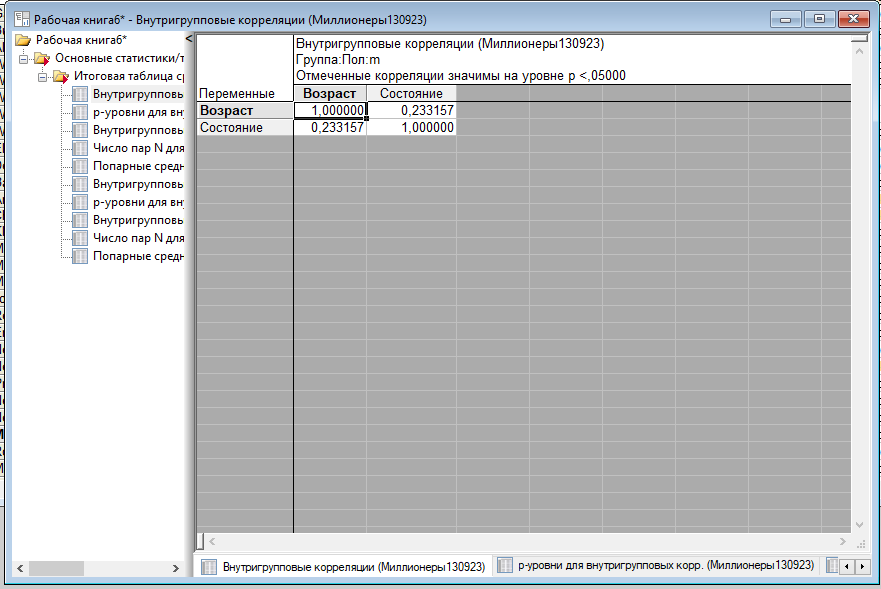
Изображение выглядит как текст, снимок экрана, График, число

Автоматически созданное описание

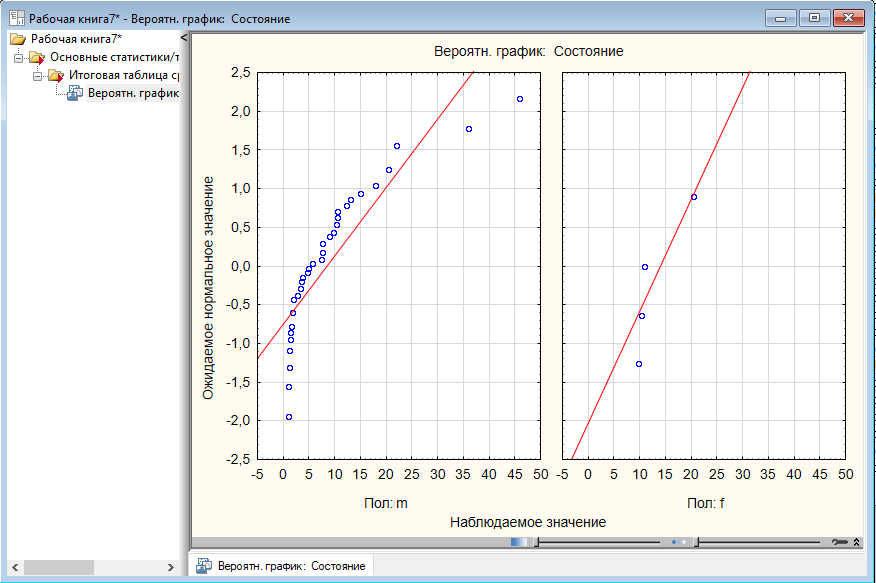
На графике видно , что средние в двух группах сильно отличаются , что можно объяснить сильным различием числа наблюдений для двух групп.



Из диаграммы видно , что средний возраст для мужчин(m) составляет 50-55 лет , в то время как для женщин(f) 60-75.



На слайде показана таблица внутригрупповой корреляции для значения m группирующей переменной .



На слайде показаны распределения значений переменной Состояние относительно доверительного интервала , в группах m и f. Из графика видно , что значений в группе f намного меньше , чем в группе m.

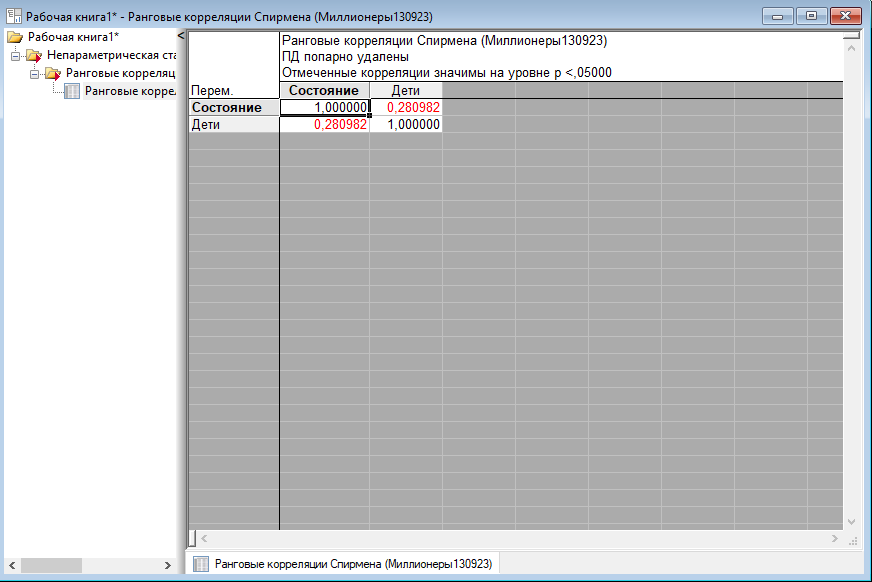
Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

На слайде показана гистограмма соответствия значений переменной Возраст в обоих группах нормальному распределению. Красной линией отмечены значения предполагаемые соответствию нормальному закону распределения. Из гистограмм видно, что значения группы f сильнее соответствуют нормальному закону, чем значения группы m. В верхней части слайда расположены формулы соотношения фактических значений групп и значений групп при нормальном распределение.

## Ранговая корреляция

Ранговая корреляция применяется в случаях , когда значения исследуемых переменных не соответствуют нормальному закону распределения ,либо объем выборки очень мал. Существуют разные методы ранговой корреляции .

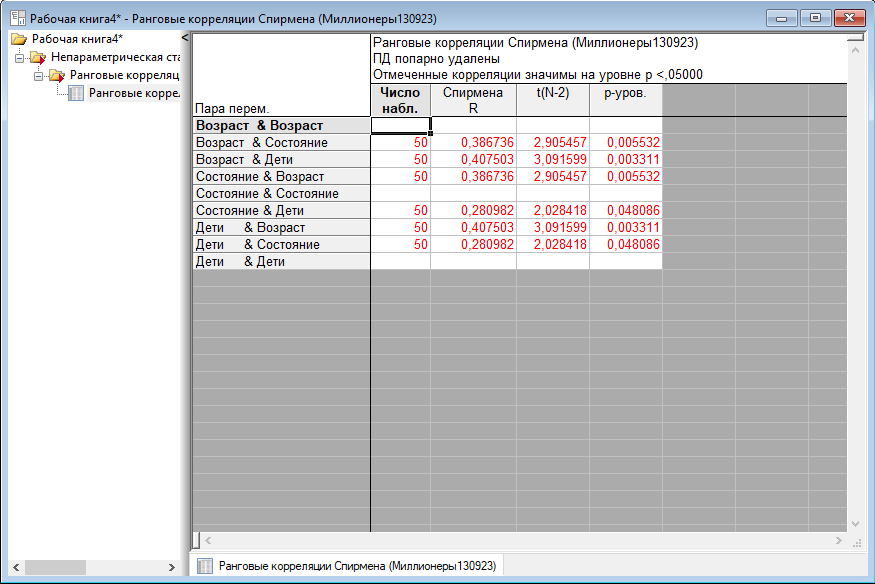


На слайде показан результат ранговой корреляции Спирмена для переменных “ Сосотояние” и “ Дети”. Данный метод предназначен для переменных , являющихся количественными . Из таблицы видно, что корреляция между переменными “Дети” и “Состояние” статистически значима , однако является близкой к слабой .

Изображение выглядит как снимок экрана, текст, программное обеспечение, дисплей

Автоматически созданное описание

На слайде приведена корреляция Спирмена для переменных “Страна”, ”Пол”, ”Социальное положение” и “Образование”.



На слайде показаны значения коэффициента спирмена , t критерия , числа наблюдений и уровня значимости для корреляции между переменными.

Существуют ,также методы для корреляции когда некоторые переменные являются категориальными.В таком случае , каждому категориальному значению такой переменной , присваивается некоторый числовой код ( ранг) и проводится корреляционный анализ с новыми значениями .

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

Корреляция Тау Кендалла используется в случаях , когда среди переменных есть переменные имеющие порядковые значения. Так как категориальные переменные могут быть переведены в порядковые , многие меж переменные корреляции являются статистически значимыми.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

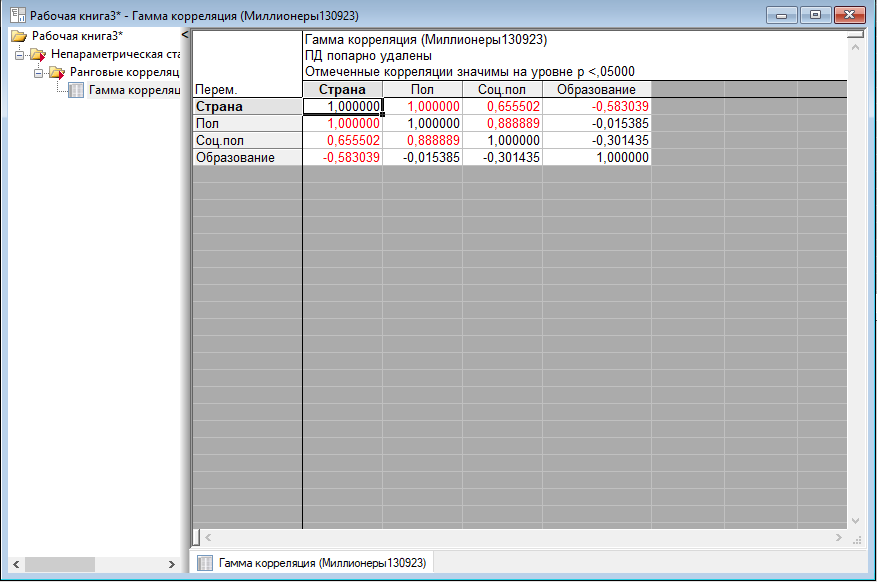
Автоматически созданное описание

На слайде показана значения коэффициента Тау-Кендалла, уровня значимости, числа наблюдений для корреляций между переменными.

Если провести данный корреляционный анализ для предыдущих переменных , то можно увидеть что корреляция между ними слабее чем корреляция Спирмена.(0,211 против 0,28 у Спирмена)

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание



Гамма корреляция применяется в случаях , когда переменные содержат много повторяющихся значений.

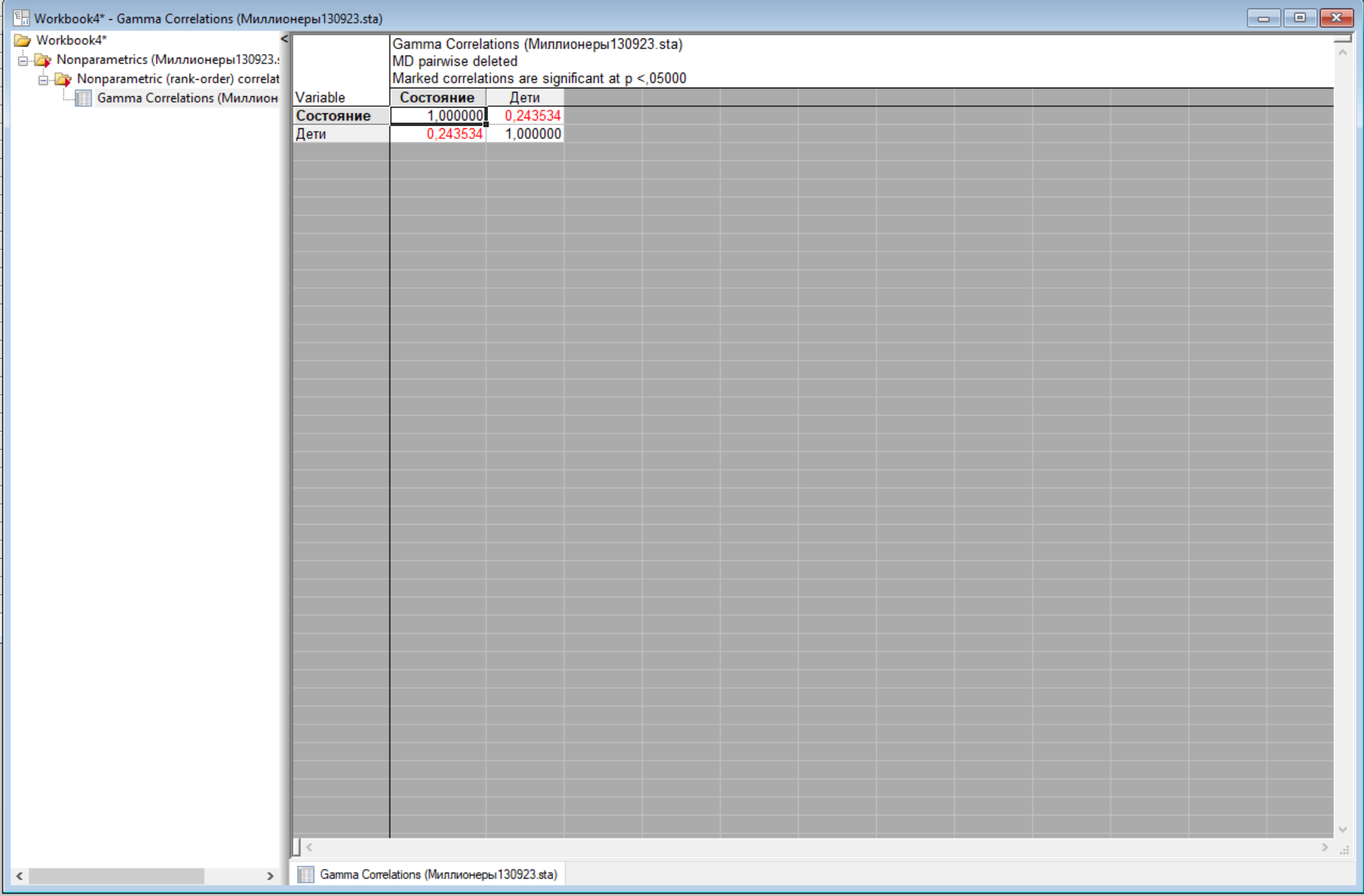
Здесь , корреляция между “Страна” и “Пол” , “Страна” и “Социальное положение”, “Страна” и “Образование” ,

“Пол” и “Социальное положение” намного сильнее случая корреляции Тау Кендалла.

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

На слайде показана значения гамма коэффициента ,уровня значимости, числа наблюдений.



Гамма Корреляция для переменных “Дети” и “Состояние” сильнее корреляции Тау Кенделла , но слабее корреляции спирмена.

Итоговая таблица значимых корреляций

|  |  |  |  |
| --- | --- | --- | --- |
|  | Спирмена | Гамма | Тау-Кендалла |
| “Страна”и”Пол” | 0,369274 | 1 | 0,3692 |
| “Страна” и ”Соц пол” | 0,283805 | 0,6555 | 0,2767 |
| “Страна” и “Образование” | -0,297249 | -0,583 | -0,2861 |
| “Пол” и “Соц пол” | 0,509984 | 0,8888 | 0,4973 |
| “Состояние” и “Дети” | 0,280982 | 0,243534 | 0,2119 |