**Отчёт об исследовании “Возраста детей из аудитории компании 7 Футов Тур”**

**Оглавление:**

[1. Собранные данные 2](#_Toc102656000)

[2. Гистограмма по данным 2](#_Toc102656001)

[3. Рассчитайте следующие значения: 2](#_Toc102656002)

[4. Данные дискретны или непрерывны? Обоснуйте. 2](#_Toc102656003)

[5. Словами опишите как строится гистограмма. 3](#_Toc102656004)

[6. Есть ли потенциальные выбросы? 3](#_Toc102656005)

[7. Определите следующие значения 3](#_Toc102656006)

[8. Постройте коробковую диаграмму 4](#_Toc102656007)

[9. Что форма диаграммы говорит о концентрации данных? 4](#_Toc102656008)

[10. Как с помощью коробковой диаграммы определить, есть ли потенциальные выбросы? 4](#_Toc102656009)

[11. Как стандартное отклонение помогает определить концентрацию данных и наличие потенциальных выбросов? 5](#_Toc102656010)

[12. Что представляет собой IQR в этой задаче? 5](#_Toc102656011)

[13. Найдите значение, которое составляет 1.5 стандартных отклонения: 5](#_Toc102656012)

[14. Собранные данные 5](#_Toc102656013)

[15. Укажите следующие значения 6](#_Toc102656014)

[16. Объясните какие выборочные характеристики можно взять в качестве точечных оценок и почему: 6](#_Toc102656015)

[17. Какие вероятностные распределения вы будете использовать при построении доверительного интервала и почему. 6](#_Toc102656016)

[18. При заданном уровне доверии a = 0.9 рассчитайте: 6](#_Toc102656017)

[19. Сколько площади в обоих хвостах в сумме y 6](#_Toc102656018)

[20. Сколько площади в каждом хвосте 6](#_Toc102656019)

[21. Изображение распределений 7](#_Toc102656020)

[22. 90% доверительный интервал содержит 90% данных? 7](#_Toc102656021)

[23. Доверительный интервал по-простому 8](#_Toc102656022)

[24. Доверительный интервал в этом исследовании 8](#_Toc102656023)

[25. Анализ доверительного оценивания. Таблица 8](#_Toc102656024)

**Исследование 1 Первичная обработка данных**

1. Собранные данные



1. Гистограмма по данным
2. Рассчитайте следующие значения:

3.1 Выборочное среднее x¯ = 9.6

3.2 Выборочное стандартное отклонение σ= s = 4.1

1. Данные дискретны или непрерывны? Обоснуйте.

Мои данные о возрасте - дискретны.

Т.к. они представлены из множества целых чисел. И продолжительность жизни человека ограничена.

1. Словами опишите как строится гистограмма.

|  |  |
| --- | --- |
| **Шаг** | **Действия** |
| 1 | Собрать данные |
| 2 | Выявить максимальное и минимальное значения данных. |
| 3 | Определить диапазон (размах) гистограммы вычитая из максимального значения данных минимальное |
| 4 | Полученный диапазон разделить на интервалы, предварительно определив их число (обычно 7-15 в зависимости от числа показателей) |
| 5 | Все данные распределить по интервалам в порядке возрастания: левая граница первого интервала должна быть меньше наименьшего из имеющихся значений |
| 6 | Подсчитать частоту каждого интервала (Сколько данных в него входит) |
| 7 | По полученным данным построить гистограмму - столбчатую диаграмму, высота столбиков которой соответствует частоте попадания данных в каждый из интервалов: |

1. Есть ли потенциальные выбросы?

Да, есть.

В таблице собранных данных умеренные выбросы выделены зеленым, а экстремальные – крсным цветом.

Подробные расчёты в отчёте Excel.

**Исследование 1.2 Исследовательский анализ данных**

1. Определите следующие значения

Характеристика Значение

Min = 2

Медиана Me = 9

Max = 27

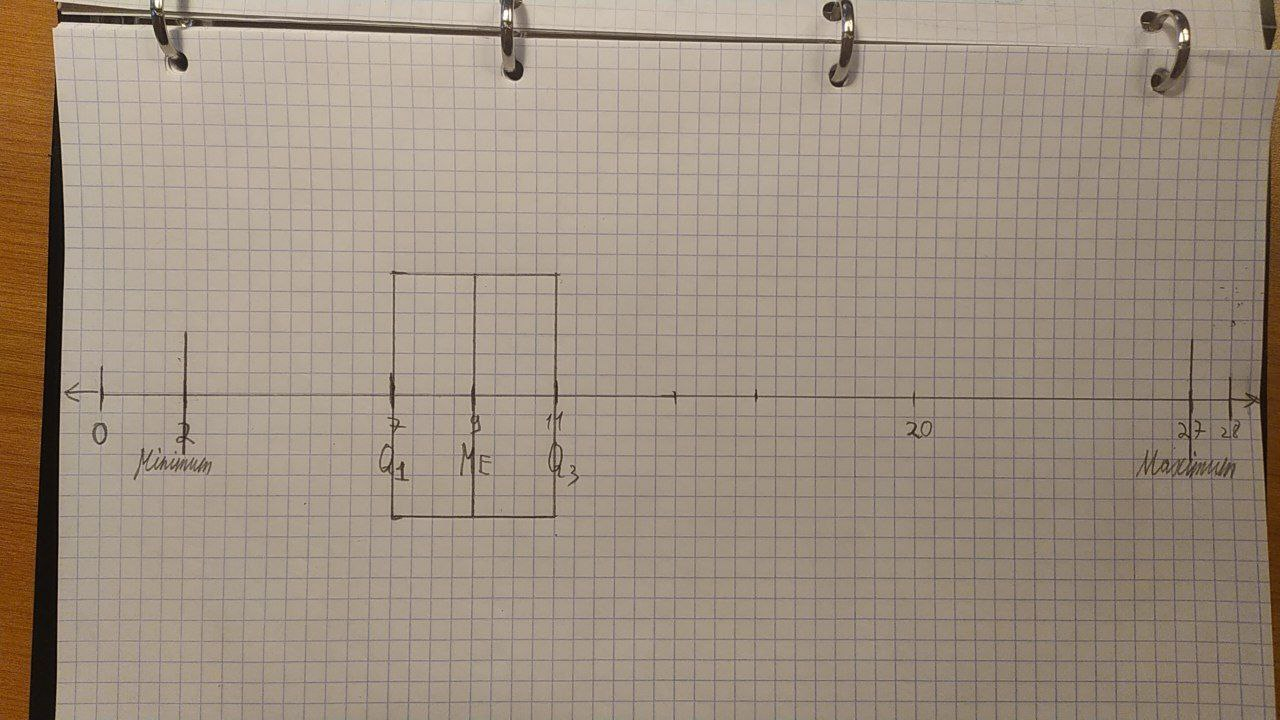
Q₁ = 7

Q₃ = 11

IQR = 4

90-й процентиль = 14

1. Постройте коробковую диаграмму



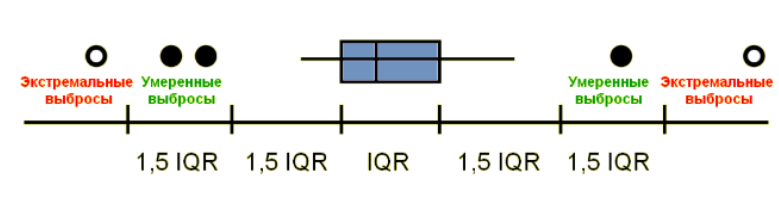
1. Что форма диаграммы говорит о концентрации данных?

Коробчатая диаграмма показывает, что наибольшая часть детей находится в возрасте от 7 до 11 лет (между 1 и 3 квартилями)

1. Как с помощью коробковой диаграммы определить, есть ли потенциальные выбросы?

Необходимо отметить на ней места, которые ограничивают диапазоны выбросов.

Если имеются данные, которые лежат в этих диапазонах, то они и являются выбросами



1. Как стандартное отклонение помогает определить концентрацию данных и наличие потенциальных выбросов?

Стандартное отклонение s определяет, насколько далеко может быть значение данных от среднего значения набора данных.

Если s маленькое, то данные сильно сконцентрированны.

Если s велико, то вероятность наличия выбросов сильно возрастает.

1. Что представляет собой IQR в этой задаче?

IQR = 4

«В моей задача IQR это разность между первыми 25 % детей и последними 25% детей по возрасту.

Т.е. он говорит, что разница между первыми 25% и 75% детей по возрасту составляет 4 года.»

1. Найдите значение, которое составляет 1.5 стандартных отклонения:

Среднее – медиана x¯ = 9.6

Стандартное отклонение s = 4.1

а. выше среднего = 15.7

b. ниже среднего = 3.5

**Исследование 2 Точечное и доверительное оценивание**

1. Собранные данные



1. Укажите следующие значения

Выборочное среднее x¯ = 9.6

Выборочное стандартное отклонение σ = s = 4.1

Объём выборки n = 95

1. Объясните какие выборочные характеристики можно взять в качестве точечных оценок и почему:

1. математического ожидания μ

Выборочное среднее x¯ - т.к. представляет из себя несмещённую оценку

2. стандартного отклонения генеральной совокупности

Выборочная дисперсия - не подходит, т.к. является смещённой оценкой.

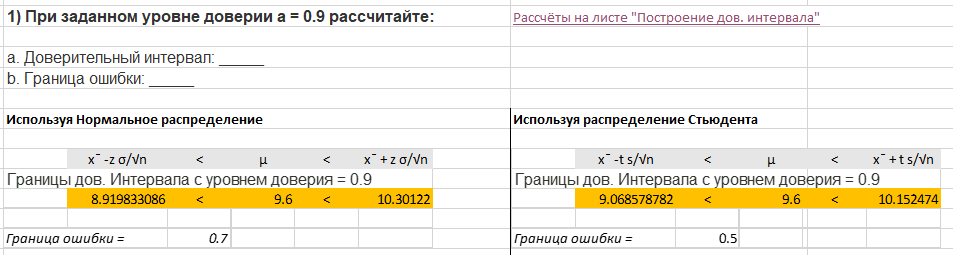
Исправленная дисперсия - подходит, т.к. является хорошей оценкой дисперсии

1. Какие вероятностные распределения вы будете использовать при построении доверительного интервала и почему.

Нормальное распределение должно дать лучшие оценки, чем распределение Стьюдента т.к. объём моей выборки n = 95, что больше 30.

*Однако даже при большом наборе данных распределение Стьюдента иногда даёт более хорошие результаты.*

1. При заданном уровне доверии a = 0.9 рассчитайте:



1. Сколько площади в обоих хвостах в сумме y

Вся площадь под графиком берётся за 1

Поэтому y = 1 - a = 1 - 0.9 = 0.1

1. Сколько площади в каждом хвосте

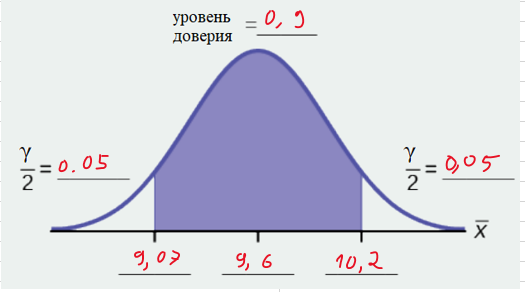
y/2 = y / 2 = 0.1 / 2 = 0.05

1. Изображение распределений

Нормальное распределение



Распределение Стьюдента



1. 90% доверительный интервал содержит 90% данных?

Доверительный интервал при Нормальном Распределении имеет границы

[8.92; 10.3]

У меня в этом промежутке находится 7 значений

Объем выборки n = 95

7 от 95 это 7.37%

Этот процент далёк от 90%

Процент данных, содержащихся в пределах доверительного интервала, может сильно отличаться от % площади, которые покрывает этот интервал.

Это связано с тем, что данные могут располагаться по-разному в рамках заданного интервала.

Их расположение зависит от распределения, по которому они размещаются.

1. Доверительный интервал по-простому

Доверительный интервал — это интервал, который можно задать для того, чтобы гарантировать что некоторый параметр лежит в заданном диапазоне.

Гарантия даётся с разным кол. процентов вероятности, поэтому размер доверительного интервала выбирается по-разному в разных ситуациях.

1. Доверительный интервал в этом исследовании

В моём исследовании доверительный интервал показывает, что скорее всего возраст ребенка в лагере будет принадлежать диапазону от 8.92 до 10.3 лет.

1. Анализ доверительного оценивания. Таблица



**Что происходит с границей ошибки по мере увеличения уровня доверия?**

По мере увеличения ур. Доверия a граница ошибки увеличивается

Чем больше ур. Доверия, тем больше данных в ключаются и тем больше размер доверительного интервала.

Т.к. граница ошибки считается, как расстояние от центра дов. Интервала, до его границы получается, что чем больше дов. Интервал и его уровень доверия, тем больше граница ошибки.

**Увеличивается или уменьшается длина доверительного интервала?**

Длина доверительного интервала увеличивается. Это происходит из-за того, что чем больше ур. Доверия, тем больше и длина интервала, т.к. покрывается больше данных.