«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Лабораторная работа №4 по теме: «Вычисление дескриптивных статистик»

Выполнил

Студент 1 курса

группы 09-115(3)

Зиновьев Е. А.

Преподаватель:

Шустова Е.П

Казань 2021

**ЗАДАНИЕ:**

1. Изучите Главу 2 "Numerical Summaries of Data" книги Rand R. Wilcox. "Understanding and applying basic statistical methods using R".
2. Выберите произвольно параметры нормального распределения.
3. Для каждого значения размера выборки n из множества {100, 500, 1000, 10 000, 100 000, 500 000, 1 000 000} вычислите следующие выборочные характеристики:
   1. выборочное среднее;
   2. усечённое среднее (trimmed mean);
   3. Винсоризованное среднее (Winsorized mean)
   4. выборочную медиану;
   5. выборочное среднее квадратичное отклонение;
   6. MADN;
   7. выборочные 25%-ый и 75%-ый квартили

и сравните их с истинными значениями оцениваемых параметров распределения. (Для наглядности представьте результаты сравнения в таблице.)

1. Сделайте выводы о состоятельности и несмещённости исследуемых оценок.

**ХОД РАБОТЫ:**

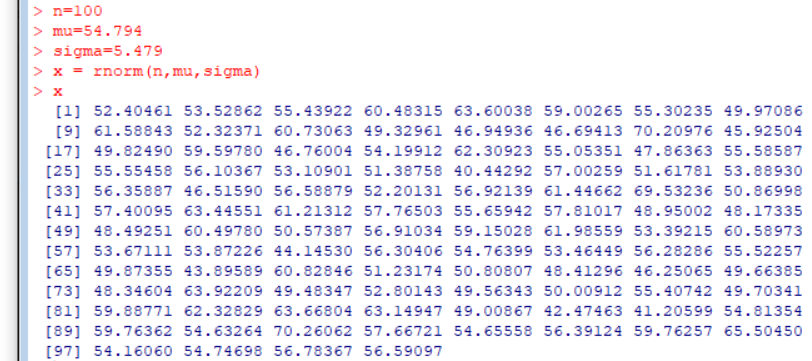
1. В качестве исходной генеральной совокупности, распределение которой можно считать близкой к нормальному: среднесуточная вероятность выпадения осадков в городе Санкт-Петербург за год

(<https://ru.wikipedia.org/wiki/Климат_Санкт-Петербурга>) - информация о среднем количестве дождливых дней

(<https://www.meteoservice.ru/weather/today/sankt-peterburg>) – информация о вероятности дождя на каждый день

2. Объем выборки: будет браться из множества {100, 500, 1000, 10 000, 100 000, 500 000, 1 000 000} , µ (среднее количество дождливых дней = 200) = 200/365\*100% = 54.794%, σ (стандартное отклонение = 20 дней) = 20/365\*100% = 5.479%.

3.1 Смоделируем нашу выборку из нормального закона распределения для объема = 100 дней (вероятность выпадения осадков в городе Санкт-Петербург при выборке в 100 дней):



Далее вычислим выборочные характеристики:

* Выборочное среднее:

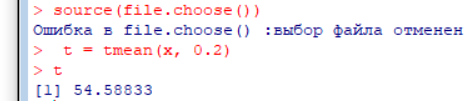


где tr – обрезка, которая используется для отбрасывания некоторых наблюдений с обоих концов отсортированного вектора (позже будет объяснено), na.rm - удаление пропущенных значений из входного вектора.

Как видим, даже при такой небольшой выборке среднее значение очень близко к среднему генеральной совокупности – разница составила 0.13%.

* Усечённое среднее (trimmed mean) с усечением 20% концов:

Для использования данной и часть последующих функций подключим пакет Rallfun, для этого используя функцию **source(file.choose()),** покажем путь к скачанному из интернета пакету.



Среднее усечённое представляет собой [статистическую](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) [меру центральной тенденции](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%80%D1%8B_%D1%86%D0%B5%D0%BD%D1%82%D1%80%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B9_%D1%82%D0%B5%D0%BD%D0%B4%D0%B5%D0%BD%D1%86%D0%B8%D0%B8), рассчитанную как среднее значение для имеющегося набора данных, из которого исключены k % наибольших и k % наименьших значений. Как правило, процент удаляемых значений устанавливается в диапазоне от 5 % до 25 %. Усечённое среднее менее чувствительно к [выбросам](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D1%80%D0%BE%D1%81_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)), чем простое [среднее арифметическое](https://ru.wikipedia.org/wiki/%D0%A1%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%B5_%D0%B0%D1%80%D0%B8%D1%84%D0%BC%D0%B5%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5), при этом оставаясь приемлемой оценкой в ряде статистических моделей. Похоже, что в данной генеральной совокупности не было сильных односторонних выбросов, так как при усечении 20% значений c каждой стороны среднее значение отклонилось от среднего генеральной совокупности чуть больше, чем прошлом пункте, а именно – 0.37%.

* Винсоризованное среднее:

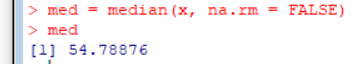
Винсоризованное среднее представляет собой [винсоризованную](https://ru.wikipedia.org/w/index.php?title=%D0%92%D0%B8%D0%BD%D1%81%D0%BE%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F&action=edit&redlink=1" \o "Винсоризация (страница отсутствует)) [статистическую](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) [меру центральной тенденции](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%80%D1%8B_%D1%86%D0%B5%D0%BD%D1%82%D1%80%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B9_%D1%82%D0%B5%D0%BD%D0%B4%D0%B5%D0%BD%D1%86%D0%B8%D0%B8) как некую конволюцию [среднего арифметического](https://ru.wikipedia.org/wiki/%D0%A1%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%B5_%D0%B0%D1%80%D0%B8%D1%84%D0%BC%D0%B5%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5) и [усеченного среднего](https://ru.wikipedia.org/wiki/%D0%A3%D1%81%D0%B5%D1%87%D0%B5%D0%BD%D0%BD%D0%BE%D0%B5_%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%B5). Расчет винсоризованного среднего сводится к тому, что k% наибольших и k% наименьших значений (обычно от 5% до 25%) заменяется наименьшими и наибольшими значениями из оставшегося массива данных, после чего рассчитывается [среднее арифметическое](https://ru.wikipedia.org/wiki/%D0%A1%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%B5_%D0%B0%D1%80%D0%B8%D1%84%D0%BC%D0%B5%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5). Винсоризованное среднее менее чувствительна к "выбросам" (аутлайерам) чем простое [среднее арифметическое](https://ru.wikipedia.org/wiki/%D0%A1%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%B5_%D0%B0%D1%80%D0%B8%D1%84%D0%BC%D0%B5%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5) при этом оставаясь приемлемой оценкой в ряде статистических моделей. Применимость винсоризованного среднего (как и [усеченного среднего](https://ru.wikipedia.org/wiki/%D0%A3%D1%81%D0%B5%D1%87%D0%B5%D0%BD%D0%BD%D0%BE%D0%B5_%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%B5)) весьма сомнительна в случаях с небольшим количеством наблюдений.



При замене крайних 20% с конца и с начала ближайшими соседями, а не просто их удалением, отклонение от среднего ген. совокупности уменьшилось – стало 0.27%, что говорит о том, что это более подходящий для данной выборки подход, чем среднее усеченное.

* Выборочная медиана:

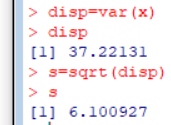
Медиана - число, которое находится в середине этого набора, если его упорядочить по возрастанию, то есть такое число, что половина из элементов набора не меньше него, а другая половина не больше. Медиана является 50-м персентилем.



Медиана лучше всех себя показала, максимально близко приблизившись к среднему ген. совокупности, – отклонение всего составило 0,0095%.

* Выборочное среднее квадратичное отклонение:

Найдем сначала сумму квадратов отклонений всех значений от среднего – дисперсию и извлечем из нее корень:



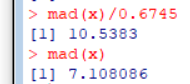
Среднеквадратическое отклонение отклонилось на 11.33% в большую сторону от среднеквадратического отклонения генеральной совокупности, что показывает более сильную изменчивость значений относительно среднего значения.

* MADN (median absolute deviation):

Cреднее абсолютное отклонение ( MAD ) -  [медиана](https://en.wikipedia.org/wiki/Median) из [абсолютных отклонений](https://en.wikipedia.org/wiki/Absolute_deviation) от медианы данных. Среднее абсолютное отклонение является мерой [статистической дисперсии](https://en.wikipedia.org/wiki/Statistical_dispersion). Более того, MAD является [надежной статистикой](https://en.wikipedia.org/wiki/Robust_statistic), более устойчивой к выбросам в наборе данных, чем [стандартное отклонение](https://en.wikipedia.org/wiki/Standard_deviation). В стандартном отклонении расстояния от [среднего](https://en.wikipedia.org/wiki/Mean) возводятся в квадрат, поэтому большие отклонения имеют больший вес, и поэтому выбросы могут сильно на него влиять. В MAD отклонения небольшого количества выбросов не имеют значения.

Для получения MADN MAD делится на N = 0.6745 (в учебнике опускается описание: “MAD is divided by 0.6745 for reasons that are difficult to explain at the moment”, но в интернете указано, что это отношение приблизительно равно стандартному отклонению: <https://en.wikipedia.org/wiki/Median_absolute_deviation>

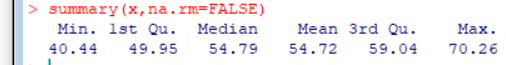
хотя судя по нижеприведенному расчету это не особо подтверждается).



Как видим, MAD оказался сильно точнее, чем он же, отнесенный к числу 0.6745: 29.77 % против 92.37 %, но менее точно, чем среднеквадратическое отклонение (11.33%).

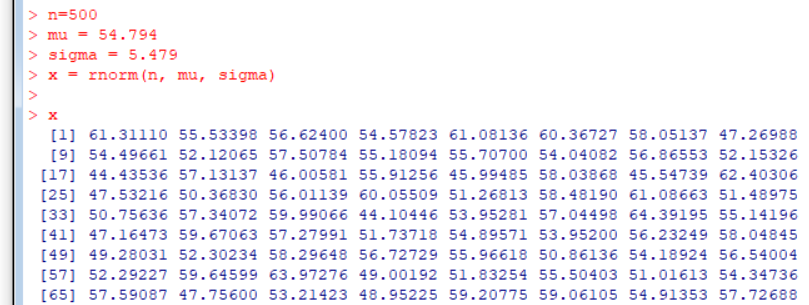
* Выборочные 25% и 75%-е квартили:

Первый квартиль (1st Qu): самые низкие 25% чисел. Третий квартиль (3rd Qu): от 51% до 75% (выше медианы).



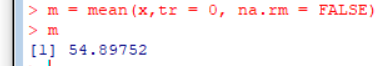
Как видим 1 и 3 квартили очень симметрично расположены относительно среднего, но левый на всего на 1 % смещен больше от среднего, то есть левый квартиль больше по размеру, чем правый (8.8% против 7.8%).

3.2 Смоделируем нашу выборку из нормального закона распределения для объема = 500 дней (вероятность выпадения осадков в городе Санкт-Петербург при выборке в 500 дней):



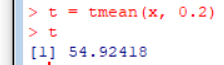
Далее вычислим выборочные характеристики:

* Выборочное среднее:



Как видим, при большей выборке среднее значение очень близко к среднему генеральной совокупности – разница составила 0.19%, что чуть больше, чем при выборке в 100 дней, но это скорее вызвано случайной генерацией при формировании нормальной выборки.

* Усечённое среднее (trimmed mean) с усечением 20% концов:



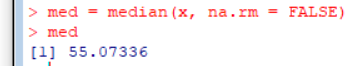
При большей выборке усеченное среднее стало ближе к среднему генеральной совокупности – разница составила 0.24% (против 0.39%).

* Винсоризованное среднее:



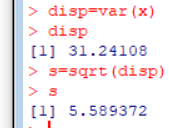
Винсоризованное среднее при большей выборке практически не изменилось – разница составила 0.29% (против 0.27%), но также более практично по сравнению с усеченной средней.

* Выборочная медиана:



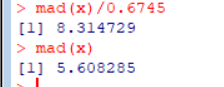
А вот медиана при большей выборке увеличила свое значение 0.5% против 0.009%, но причина опять же связана со случайной генерацией при формировании нормальной выборки.

* Выборочное среднее квадратичное отклонение:



Среднеквадратическое отклонение отклонилось на 2% в большую сторону от среднеквадратического отклонения генеральной совокупности, что показывает более сильную изменчивость значений относительно среднего значения, но зато более слабую против выборки в 100 дней (11.33%).

* MADN (median absolute deviation):

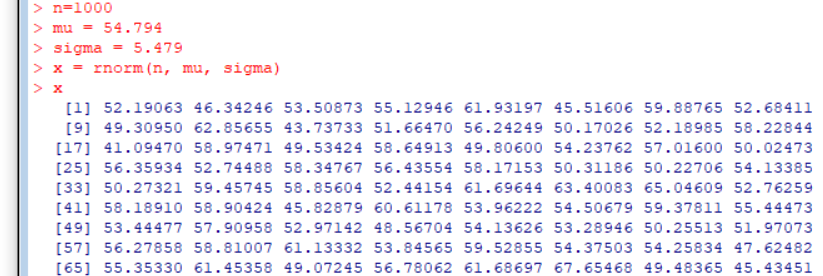


Как видим, теперь MAD оказался менее точным, чем он же, отнесенный к числу 0.6745: против 51.74 % против 2 %.

* Выборочные 25% и 75%-е квартили:

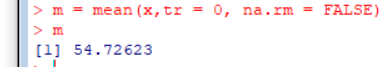
При данной сформированной выборке квартили теперь почти абсолютно симметричны (вплоть до сотых значений), а также диапазоны стали уже (6.9 % отклонение от среднего против 8.84%)

3.3 Смоделируем нашу выборку из нормального закона распределения для объема = 1000 дней (вероятность выпадения осадков в городе Санкт-Петербург при выборке в 1000 дней):



Далее вычислим выборочные характеристики:

* Выборочное среднее:



Как видим, при большей выборке среднее значение очень близко к среднему генеральной совокупности – разница составила 0.12%, но в этот раз меньше, чем при выборке в 100 дней и 500 дней.

* Усечённое среднее (trimmed mean) с усечением 20% концов:



При большей выборке усеченное среднее стало ближе к среднему генеральной совокупности – разница составила 0.11% (против 0.39% для выборки в 100 дней).

* Винсоризованное среднее:



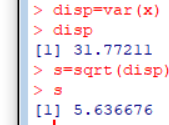
Винсоризованное среднее при большей выборке тоже практически не изменилось – разница составила 0.28% (против 0.29%), но в этот раз менее практично по сравнению с усеченной средней.

* Выборочная медиана:



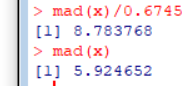
Медиана при большей выборке также уменьшила свое значение 0.01% против 0.5%.

* Выборочное среднее квадратичное отклонение:



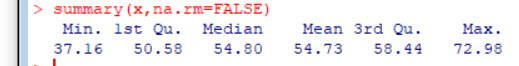
Среднеквадратическое отклонение слегка увеличилось относительно выборки в 500 дней, но зато все равно сильно меньше выборки в 100 дней (11.33%).

* MADN (median absolute deviation):



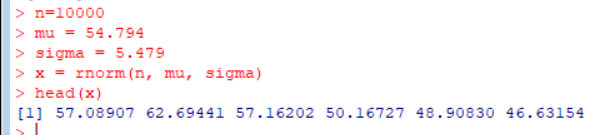
Как видим, теперь MAD оказался более точным, чем он же, отнесенный к числу 0.6745: 8,04 % против 60 % и даже точнее, чем среднеквадратическое отклонение (11.33%).

* Выборочные 25% и 75%-е квартили:



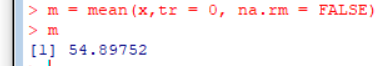
Как видим 1 и 3 квартили очень симметрично расположены относительно среднего (7.6% против 6.6%), но левый на всего на 1 % смещен больше от среднего, то есть левый квартиль больше по размеру, чем правый.

3.4 Смоделируем нашу выборку из нормального закона распределения для объема = 10 000 дней (вероятность выпадения осадков в городе Санкт-Петербург при выборке в 10 000 дней):



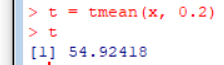
Далее вычислим выборочные характеристики:

* Выборочное среднее:



Как видим, при большей выборке среднее значение очень близко к среднему генеральной совокупности – разница составила 0.19%, что чуть больше, чем при выборке в 100 дней, но это скорее вызвано случайной генерацией при формировании нормальной выборки.

* Усечённое среднее (trimmed mean) с усечением 20% концов:



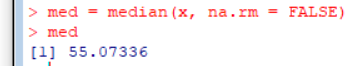
При большей выборке усеченное среднее стало ближе к среднему генеральной совокупности – разница составила 0.24% (против 0.39%).

* Винсоризованное среднее:



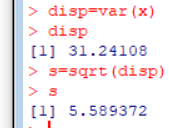
Винсоризованное среднее при большей выборке практически не изменилось – разница составила 0.29% (против 0.27%), но также более практично по сравнению с усеченной средней.

* Выборочная медиана:



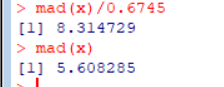
А вот медиана при большей выборке увеличила свое значение 0.5% против 0.009%, но причина опять же связана со случайной генерацией при формировании нормальной выборки.

* Выборочное среднее квадратичное отклонение:



Среднеквадратическое отклонение отклонилось на 2% в большую сторону от среднеквадратического отклонения генеральной совокупности, что показывает более сильную изменчивость значений относительно среднего значения, но зато более слабую против выборки в 100 дней (11.33%).

* MADN (median absolute deviation):

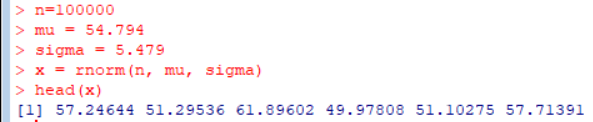


Как видим, теперь MAD оказался менее точным, чем он же, отнесенный к числу 0.6745: против 51.74 % против 2 %.

* Выборочные 25% и 75%-е квартили:

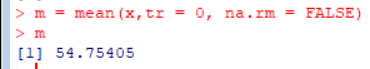
При данной сформированной выборке квартили теперь почти абсолютно симметричны (вплоть до сотых значений), а также диапазоны стали уже (6.9 % отклонение от среднего против 8.84%)

3.5 Смоделируем нашу выборку из нормального закона распределения для объема = 100 000 дней (вероятность выпадения осадков в городе Санкт-Петербург при выборке в 100 000 дней):



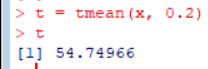
Далее вычислим выборочные характеристики:

* Выборочное среднее:



Отклонение среднего выборки от среднего ген.совокупности уменьшилось до 0.08% (против меньших выборок 0.17%).

* Усечённое среднее (trimmed mean) с усечением 20% концов:



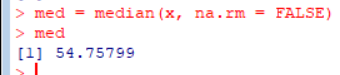
Усеченное среднее до тысячных совпало с выборочным средним – 0.08% (против 0.25%).

* Винсоризованное среднее:



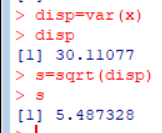
Аналогично и с винсоризованным средним, тоже отклонение составило– 0.08% (против 0.28% прошлых выборок).

* Выборочная медиана:



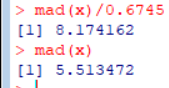
Аналогично и с выборочной медианой, тоже отклонение уменьшилось – 0.065% (против 0.17% прошлых выборок).

* Выборочное среднее квадратичное отклонение:



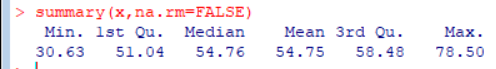
Отклонение среднеквадратического отклонения от среднкв. отклонения ген.совокупности еще более сильно уменьшилось относительно меньших выборок (0.14% против 2.5%).

* MADN (median absolute deviation):



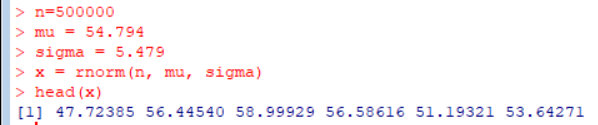
Как видим, теперь MAD оказался более точным, чем он же, отнесенный к числу 0.6745: 0.56 % против 49 %.

* Выборочные 25% и 75%-е квартили:



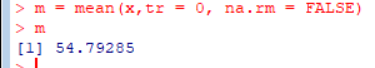
Как видим 1 и 3 квартили очень симметрично расположены относительно среднего (6.8% против 6.7%).

3.6 Смоделируем нашу выборку из нормального закона распределения для объема = 500 000 дней (вероятность выпадения осадков в городе Санкт-Петербург при выборке в 500 000 дней):



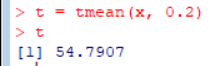
Далее вычислим выборочные характеристики:

* Выборочное среднее:



Отклонение среднего выборки от среднего ген.совокупности уменьшилось до 0.004% (против меньших выборок 0.13%, прошлая выборка – 0.08%).

* Усечённое среднее (trimmed mean) с усечением 20% концов:



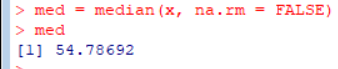
Отклонение усеченного среднего аналогично сильно уменьшилось – 0.0073% (против прошлой выборки 0.08%).

* Винсоризованное среднее:



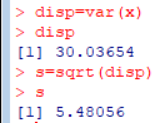
Аналогично и с винсоризованным средним, тоже отклонение сильно уменьшилось и составило– 0.0018% (против 0.2% прошлых выборок).

* Выборочная медиана:



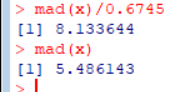
Аналогично и с выборочной медианой, тоже отклонение сильно уменьшилось – 0.012% (против 0.13% прошлых выборок).

* Выборочное среднее квадратичное отклонение:



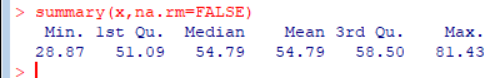
Отклонение среднеквадратического отклонения от среднкв. отклонения ген.совокупности еще более сильно уменьшилось относительно меньших выборок (0.018% против 0.14% при последней выборки).

* MADN (median absolute deviation):



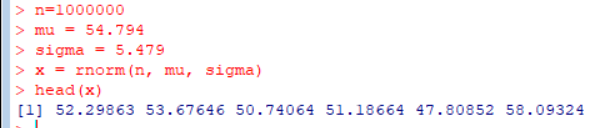
Как видим, теперь MAD оказался более точным, чем он же, отнесенный к числу 0.6745: 0.12 % против 48 %.

* Выборочные 25% и 75%-е квартили:



Как видим 1 и 3 квартили очень симметрично расположены относительно среднего (6.7% против 6.7%).

3.7 Смоделируем нашу выборку из нормального закона распределения для объема = 1 000 000 дней (вероятность выпадения осадков в городе Санкт-Петербург при выборке в 1 000 000 дней):



Далее вычислим выборочные характеристики:

* Выборочное среднее:



Отклонение среднего выборки от среднего ген.совокупности увеличилось до 0.0073% (против прошлой выборки 0,03%), что связано скорее с небольшим отклонением при случайной генерации чисел, что немного ухудшило качество выборки.

* Усечённое среднее (trimmed mean) с усечением 20% концов:



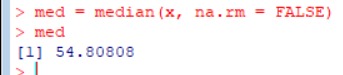
Аналогично можно сказать и про отклонение усеченного среднего, оно увеличилось – 0.01% (против прошлой выборки 0.0073%).

* Винсоризованное среднее:



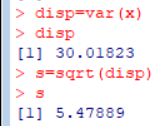
Аналогично и с винсоризованным средним, тоже отклонение немного увеличилось и составило– 0.009% (против 0.0018% прошлой выборки).

* Выборочная медиана:



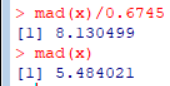
Аналогично и с выборочной медианой, тоже отклонение немного увеличилось – 0.025% (против 0.012% прошлой выборки).

* Выборочное среднее квадратичное отклонение:



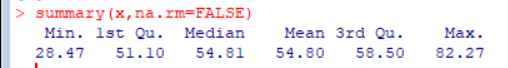
А вот отклонение среднеквадратического отклонения от среднкв. отклонения прошлой выборки почти не изменилось, совпало в пределах точности.

* MADN (median absolute deviation):



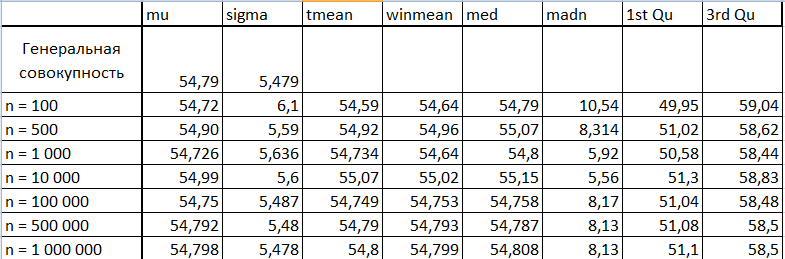
Как видим, теперь MAD оказался более точным, чем он же, отнесенный к числу 0.6745: 0.018 % против 48 %.

* Выборочные 25% и 75%-е квартили:

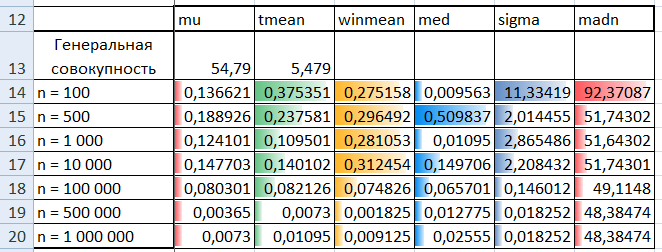


Как видим 1 и 3 квартили очень симметрично расположены относительно среднего (6.7% против 6.7%), аналогично прошлой выборке.

Итоговая таблица выборочных характеристик вероятностей выпадения осадков для каждой выборки:



Итоговая таблица отклонения выборочных характеристик от значений генеральной совокупности (отклонение в процентном соотношении):



4. Мы видим, что с увеличением объема выборки выборочные характеристики в целом становятся все ближе к истинным значениям, но иногда в больших выборках появляются отклонения, слегка превышающие отклонения в меньших выборках, но это скорее связано со случайной генерацией чисел, немного снизившей качество выборки. И вот по этой причине с одной стороны, если взять 2 ближайшие выборки, отклонение может, как увеличиться, так и уменьшиться, а с другой, если хотя бы через одну выборку, то почти всегда уменьшаются. Но так как тенденция идет на уменьшение отклонения от истинных значений, то есть приближение к мат. ожиданию данных параметров, то данные выборочные характеристики можно назвать состоятельными.

Выборочный показатель, который при многократном повторении выборки стремится к своему теоретическому значению, называется несмещенной оценкой. Выборочная средняя – это несмещенная оценка математического ожидания, так как средняя из выборочных средних стремится к своему теоретическому значению по генеральной совокупности.

Дисперсия может быть, как смещенной, так и несмещенной. Соотношение между смещенной и несмещенной дисперсией составляет n/n-1. Несложно догадаться, что с ростом n (объема выборки) данное выражение стремится к 1, то есть разница между значениями смещенной и несмещенной дисперсией уменьшается, а при переходе к среднеквадратичному отклонению по выборке (корень из выборочной дисперсии) разница становится еще меньше. Таким образом, если в небольших выборках эффект смещенной дисперсии может проявиться хоть немного, то в нашем случае, это слабо отражено, поэтому и среднеквадратическое отклонение является несмещенной характеристикой.

Усеченное среднее и винсоризованное среднее также дают несмещенную оценку при нормальном распределении, поэтому все выборочные характеристики можно назвать состоятельными и несмещенными, несмотря на небольшие отклонения в большую сторону при более широких выборках из-за псевдослучайной генерации чисел.