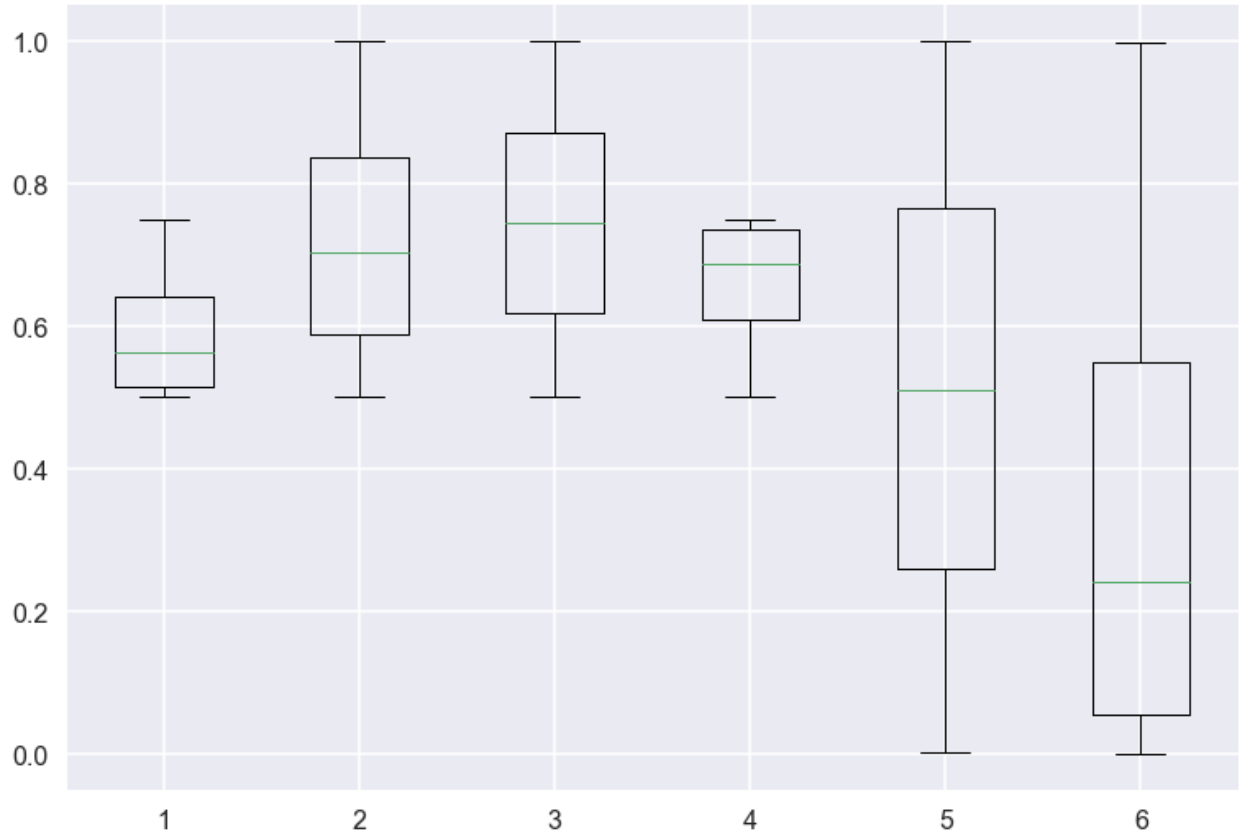
1. Смотрим на распределения (80 баллов)

(i). У вас есть массив значений, равномерно распределенных между 0 и 1 (т. е. вероятность, что значение будет принадлежать конкретному отрезку, например [0.5, 0.6], зависит только от длины отрезка).

К этому массиву были применены следующие функции, чтобы получить новые массивы:

A. $f(x) = \ -x^2 + x + 0.5$
B. $f(x) = \ 1 - x$
C. $f(x) = 2^{x-1} = \frac{2^x}{2}$
D. $f(x) = \ x^2$
E. $f(x) = \frac{x}{2} + 0.5$



Выше приведены ящики с усами, которые были построены на полученных массивах. Вам нужно найти соответствие между функциями и их визуализацией и обосновать свой ответ.

Решение:

Рассмотрим функцию A. Ее минимальное и максимальное значение на отрезке [0, 1] равны 0.5 и 0.75. Значит, соответствующий ящик – либо 1, либо 4. Давайте найдем медиану для A. Для этого мы посчитаем половину максимальных значений. Половина значений в начальной выборке находится между 0.25 и 0.75, то есть после преобразования эти

значения будут от 0.6875. При этом значение функции до 0.25 и после 0.75 меньше 0.6875. Значит, 0.6875 является медианной. То есть, A – 4 график.

Заметим, что при монотонном преобразовании (т.е. применении убывающей или возрастающей функции) медианное наблюдение останется медианным, т.к. оно все еще больше половины наблюдений. Медиана изначального наблюдения равна 0.5. Применим функции к 0.5 и получим:

A.
B. 0.5
C. 0.7
D. 0.25
E. 0.75

Значит, B – 5. D – 6.

Осталось распределить C и E. Заметим, что E – линейная функция. Значит, ее ящик с усами будет подобен изначальному ящику с усами, то есть примерно симметричный относительно медианы. Значит, E – 3. А оставшейся C соответствует второй график, так как медиана должна быть равна 0.7, а максимум – единице.

Получаем

A. 4
B. 5
C. 2
D. 6
E. 3

Критерии:

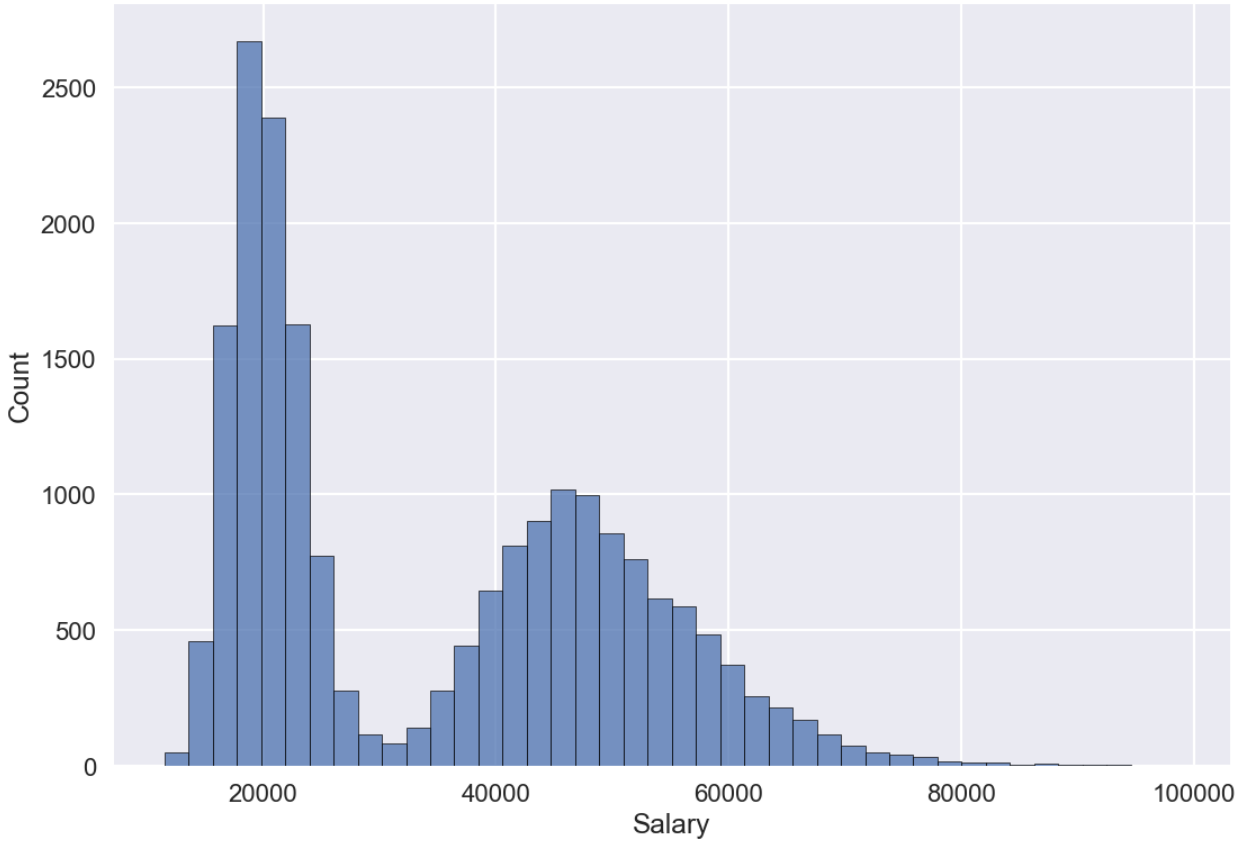
Max – **44 балла**

По 4 балла за каждое верное соответствие

8 баллов за верное объяснение соответствия A

По 4 балла за каждое верное объяснение соответствия B-E

(ii). Вы проводите исследование зарплат сотрудников некой компании. Для этого вы провели опрос, разослав всем электронное письмо. По результатам опроса вы построили распределение зарплат, которое можно увидеть ниже:



Также вы заполнили таблицу.

Средняя зарплата	34 839
Медианная зарплата	30 622
75-й персентиль	48 272
25-й персентиль	19 979
Максимальная зарплата	98 710

Но руководитель HR-департамента компании, у которого уже есть все данные о зарплатах, по секрету рассказал вам, что опрос прошли почти все сотрудники с зарплатой выше 33 тыс. рублей, но из остальных приняла участие только половина.

Оцените, какие из значений в вашей таблице завышены, а какие занижены, и почему.

Опишите, как истинное распределение зарплат будет отличаться от представленного выше.

Решение:

У нас отсутствует половина наблюдений с зарплатой ниже 33 тысяч рублей, для несмещенной оценки нам нужно их доопросить. Т.е. в выборку добавятся наблюдения до 33

тысяч рублей. Значит, средняя зарплата и 75% персентиль уменьшатся, так как они больше, чем добавляемые значения. Медианная зарплата может как измениться в любую сторону, так и не измениться, все зависит от медианного значения новых наблюдений. Если мы предположим, что вероятность ответа среди людей с низкой зарплатой не зависит от зарплаты (т.е. у всех равно 0.5), то медиана добавляемых значений будет в районе 20тыс и общая медиана уменьшится. Аналогичные рассуждения для 25% персентиля. Максимальная зарплата является оценкой снизу, так как самый высокооплачиваемый сотрудник мог не пройти опрос.

В истинном распределении зарплат часть левее 33 тысяч рублей будет выше. При равномерном распределении вероятности ответа среди людей с низкой зарплатой левый «колокольчик» будет в 2 раза выше.

Критерии:

Max – **36 баллов**

Средняя зарплата завышена – **4 балла** (**1 балл** при отсутствии обоснования)
75-й персентиль завышен – **4 балла** (**1 балл** при отсутствии обоснования)
Медиана – (неизвестно + обоснование) или (завышена + предположение о равномерном распределении вероятности ответа) – **8 баллов**. Если предположение использовано, но не обговорено отдельно – **4 балла**.
25-й персентиль – (неизвестно + обоснование) или (завышена + предположение о равномерном распределении вероятности ответа) – **8** баллов. Если предположение использовано, но не обговорено отдельно – **4** балла.
Максимум – **4** балла (**2 балла** при отсутствии обоснования)

За вторую часть максимум 8 баллов:
Если в решении использовано предположение, что вероятность ответа распределена равномерно, и обосновано удвоение левой части – **8** баллов
Если предположение не использовано и обосновано увеличение (нет части про удвоение или пропорциональный сдвиг) – **8** баллов
Если предположение использовано и обосновано увеличение (нет части про удвоение или пропорциональный сдвиг) – **4** балла
Если предположение не использовано и обосновано удвоение – **4** балла
Ответ без обоснования – **1** балл