

Время выполнения задания – 180 минут. Максимальное количество баллов – 100

Задание 1. (15 баллов) Юный аналитик Андрей получил заказ от правительства Свердловской области: проанализировать зависимость наличия личного транспорта от пола среди жителей Свердловской области.

1. (7 баллов) Сформулируйте потенциальную *гипотезу**, которую Андрей мог проверять в ходе своего исследования.

2. (8 баллов) Сформулируйте экономический *механизм**, который оправдывает гипотезу из пункта 1.

*Гипотеза исследования** - ваше предположение о наличии и характере связи между конкретными факторами (переменными), взаимосвязь которых является темой исследования.

*Механизм гипотезы** - подробное обоснование правильности выдвинутой гипотезы.

Решение и критерии оценки

1. Гипотеза должна быть построена о взаимосвязи двух факторов, один из которых отражает пол, а другой - наличие личного транспорта. Пример верной гипотезы: в Свердловской области женщины с более высокой вероятностью владеют личным транспортом. Участник получает 7 баллов за верную гипотезу. Участник получает 0 баллов за ошибки в построении гипотезы (например, если факторы не связаны с переменными, упомянутыми в задании, или количество факторов не равно двум; если отсутствует предположение о направлении связи, не упомянута Свердловская область).
2. Пример верного механизма: женщины, в среднем, чаще берут на себя бытовые дела, поэтому им приходится чаще передвигаться по городу, из-за чего они склонны иметь личный транспорт с более высокой вероятностью, чем мужчины. Участник получает 8 баллов за верный механизм. Участник получает 0 баллов, если механизм построен неверно или недостаточно подробно, а также если он не связан с верной гипотезой из предыдущего пункта, или если за гипотезу в пункте 1 было получено 0 баллов.

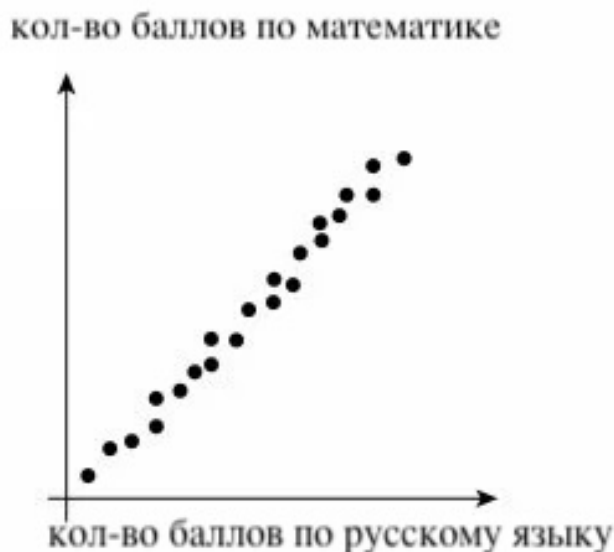
Задание 2. (15 баллов) Школьница Оля получила от местного университета массив данных в Excel, который содержит информацию об абитуриентах, поступивших в него в текущем году. Массив содержит следующие переменные (переменные указаны в том порядке, в котором расположены в данных Оли): порядковый номер абитуриента, имя, фамилия, город, возраст, пол, количество баллов ЕГЭ по математике, количество баллов ЕГЭ по русскому языку, количество баллов ЕГЭ по биологии. Для проверки основной гипотезы исследования Оля написала следующую команду:

KOPPEЛ(G2:G1000, H2:H1000)

1. (3 балла) Какие две переменные стали фокусом исследования, которое проводила Оля?
2. (3 балла) Какое количество абитуриентов попало в выборку исследования?
3. (3 балла) Предположим, что в результате написания команды, Оля получила значение 0.9. Какой вывод можно сделать о взаимосвязи переменных из пункта 1 (о направлении этой взаимосвязи и ее силе)?
4. (3 балла) Изобразите, как выглядит диаграмма рассеивания для данной взаимосвязи.
5. (3 балла) Назовите два подходящих типа графика для того, чтобы показать распределение переменной «возраст». Объясните свой выбор (не более двух предложений).

Решение и критерии оценки

1. Лист в Excel начинается с буквы А и далее идет в порядке букв английского алфавита. Буквы G и I занимают 7-е и 8-е места соответственно. Переменные, находящиеся в фокусе исследования, - это количество баллов ЕГЭ по математике и количество баллов ЕГЭ по русскому языку. Участник получает 3 балла, если обе переменные указаны верно. Участник получает 0 баллов, если хотя бы одна из переменных указана неверно.
2. Команда показывает, что корреляция строилась со второго по 1000-е наблюдение (поскольку, как правило, в первой строке хранятся названия переменных). Количество наблюдений равно $1000 - 2 + 1 = 999$. Участник получает 3 балла за верный ответ и 0 - за неверный.
3. Значение корреляции 0.9 является положительным числом, которое очень близко к максимальному значению корреляции (1). Из этого следует, что корреляция между переменными сильная и положительная. Участник получает 3 балла за правильный ответ. Участник получает 0 баллов, если не сделан вывод о положительной корреляции или ее силе.
4. Правильным графиком будут графики наподобие тех, что представлены на картинке, или иные вариации диаграммы рассеивания с положительной корреляцией. Участник получает 3 балла за верный график. Участник получает 0 баллов за другой тип графика, изображение нулевой/отрицательной или слабой корреляции, или если оси не подписаны/подписаны неверно.



5. Пол является качественной переменной, для которой подойдут, например, круговая и столбчатая диаграммы. Участник получает 3 балла за верные типы графиков с обоснованием. Типы графиков, подходящие только для непрерывных переменных, а также несуществующие типы графиков, не являются верными. Участник получает 0 баллов за отсутствие объяснения.

Задание 3. (14 баллов) Аптека «ТюменьЗдрав» испытывает спад выручки за последние полтора месяца. Руководство компании проанализировало данные и выяснило, что группа покупателей, которая стала меньше посещать аптеку - это пенсионеры. Вас наняли для исследования вопроса: «Что влияет на выбор аптеки среди пенсионеров Тюмени?» и попросили провести исследование методом анкетирования, чтобы получить как можно более подробные ответы.

Чтобы не опрашивать всю генеральную совокупность (все пенсионеры Тюмени) и сократить расходы на исследование, руководство аптеки предложило несколько вариантов формирования выборки для анкетирования. Объясните, почему каждая из предложенных выборок не является *репрезентативной** (не более одного предложения на пункт).

1. (2 балла) Пенсионеры Тюмени, которые регулярно посещают «ТюменьЗдрав».
2. (2 балла) Пенсионеры Тюмени, которые никогда не посещали «ТюменьЗдрав».
3. (2 балла) 50000 случайно отобранных пенсионеров Свердловской области.
4. (2 балла) Пенсионеры Тюмени, проживающие в Ленинском административном районе (где и расположена аптека «ТюменьЗдрав»).
5. (2 балла) Пенсионеры Тюмени, которые имеют проблемы со здоровьем.
6. (2 балла) Врачи поликлиник Тюмени, где обслуживаются пенсионеры.
7. (2 балла) Дети и родственники пенсионеров Тюмени, которые занимаются покупкой лекарств.

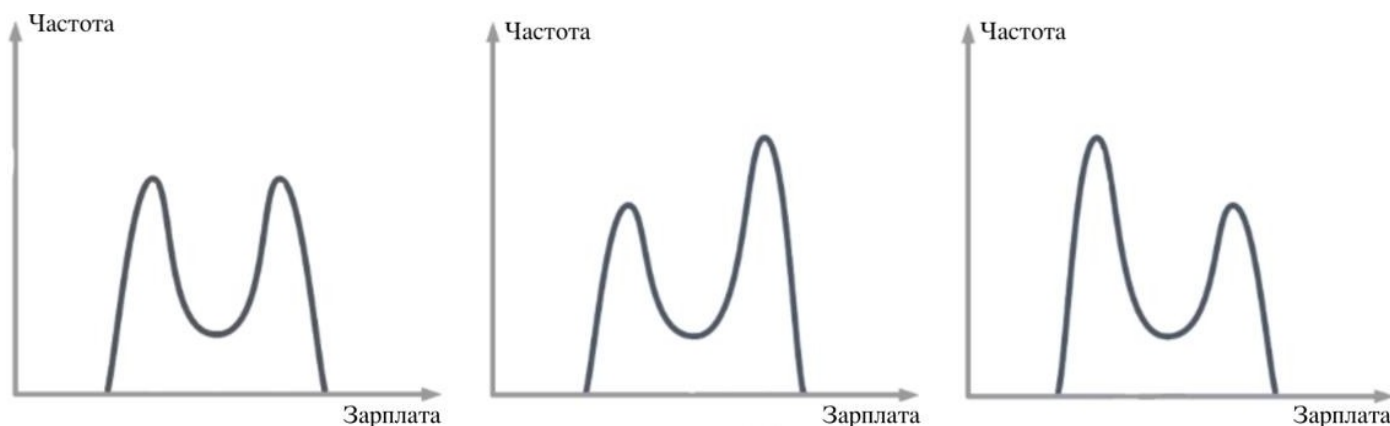
*Репрезентативная выборка** -- выборка, свойства которой совпадают со свойствами генеральной совокупности (в ней представлены все подгруппы, которые входят в генеральную совокупность).

Решение и критерии оценки

В ряде примеров произошла подмена генеральной совокупности пенсионеров Тюмени на других людей (пункты 4, 7, 8), поэтому выборка не может быть репрезентативной. В остальных случаях (пункты 1, 2, 5, 6) генеральная совокупность определена верно, но способ формирования выборки приводит к смещению. В ходе формирования выборки наша задача - сделать отбор как можно более случайным и беспристрастным, не основывая его на конкретных факторах. Например, фокусируясь на людях из конкретного района, мы можем оказаться в ситуации, когда опросим более богатых людей, чем в целом по Тюмени. Участник получает по 2 балла за каждый верный ответ и 0 - если ответ неверный.

Задание 4. (16 баллов) Долгое время в компании «Альфа» распределение зарплат сотрудников совпадало с нормальным.

1. (5 баллов) Изобразите данное распределение на графике схематично, указав на нем: среднее значение, моду, медиану.
2. (5 баллов) Аналитики «Альфа» заметили, что в последнем месяце три показателя находятся в следующем соотношении: среднее значение > медиана > мода. Изобразите данное распределение на графике схематично, указав на нем: среднее значение, моду, медиану.
3. (6 баллов) Аналитики «Альфа» связались с тремя компаниями-партнерами «Бета», «Гамма», «Омега» и запросили графики распределения зарплат среди их сотрудников. На каждом графике укажите: среднее значение, медиану, моду.



Решение и критерии оценки

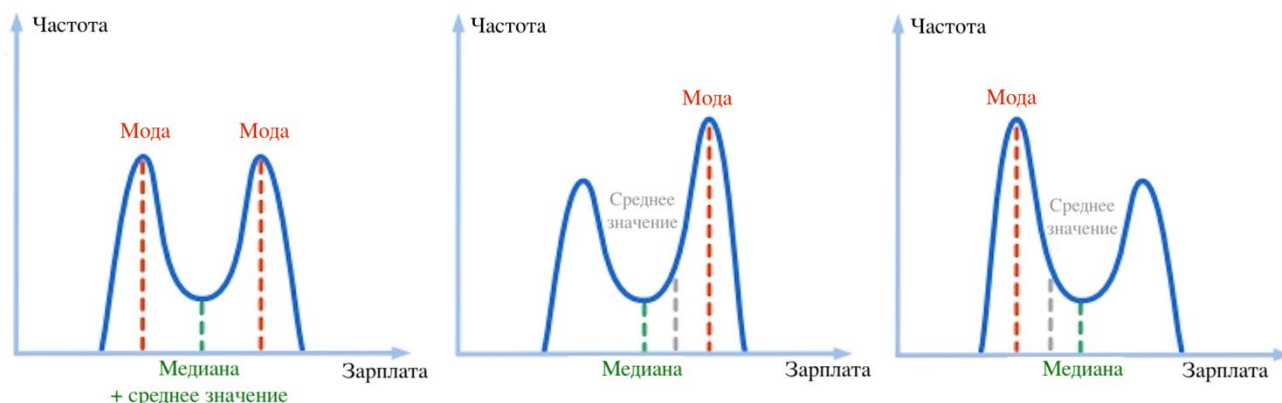
1. Участник получает 5 баллов за верный график. Участник получает 0 баллов за неверное изображение распределения, неверно указанные точки, неподписанные/неверно подписанные оси.



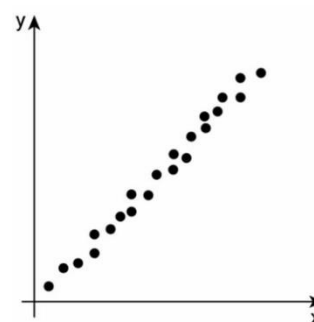
2. Участник получает 5 баллов за верный график. Участник получает 0 баллов за неверное изображение распределения, неверно указанные точки, неподписанные/неверно подписанные оси.



3. Участник получает по 2 балла за верный график, 0 - за неверное изображение распределения, неверно указанные точки, неподписанные/неверно подписанные оси.



Задание 5. (15 баллов) Отличник Паша работал над проектом по анализу данных и исследовал взаимосвязь между доходами человека и количеством совершенных авиаперелетов. Собрав репрезентативные данные по своему городу и построив диаграмму рассеивания, Паша увидел взаимосвязь, представленную на графике. Он вспомнил, что на лекции по анализу данных им рассказывали, что такой график показывает сильную положительную корреляцию, и сделал вывод про свой город: «Увеличение доходов человека положительно влияет на количество авиаперелетов». За такой вывод Паша получил двойку и перестал быть отличником.



- (5 баллов) Определите, в чем состоит ошибка, и как называется феномен, который не смог распознать Паша. Почему он имеет место быть в данном примере?
- (5 баллов) Исправьте ошибку Паши и сформулируйте корректный вывод.
- (5 баллов) Предложите, как можно было бы изменить исследование, чтобы с выводом Паши нельзя было поспорить.

Решение и критерии оценки

- Феномен, имеющий место в данном примере, называется «эндогенность». Идея заключается в том, что между факторами невозможно установить причинно-следственную взаимосвязь по ряду причин. Например, здесь может быть обратная взаимосвязь, при которой количество авиаперелетов ведет к увеличению доходов, если работа человека связана с командировками, за которые он получает дополнительный доход. Другой причиной эндогенности может быть влияние общего фактора, такого как возраст. С увеличением возраста человек может больше зарабатывать благодаря накопленному опыту работы, а также совершать большее количество перелетов в результате расширения кругозора и интереса к путешествиям. Участник получает 5 баллов за объяснение невозможности установления причинно-следственной связи.

2. Корректный вывод должен говорить не о причинно-следственной связи, а о простой корреляции, которая наблюдается на графике: доходы человека и количество авиаперелетов положительно коррелируют. Участник получает 5 баллов за вывод о корреляции.
3. Чтобы вывод о причинно-следственной взаимосвязи оказался верным, графика корреляции недостаточно. Нужно усложнить исследование, включив в него другие факторы, которые связаны с обеими переменными: возраст, место жительства, пол и другие. Однако даже этого будет недостаточно для борьбы с эндогенностью: необходимо создать случайное изменение в доходах, например, предоставить случайным людям определенную сумму. Далее мы сравниваем их количество авиаперелетов с похожими людьми, которые не получили деньги. Участник получает 5 баллов за предложение способа борьбы с эндогенностью.

Задание 6. (25 баллов) Предложите дизайн собственного исследования на интересную для вас тему и раскройте его как можно подробнее, упомянув:

1. (5 баллов) Факторы, которые являются фокусом исследования, гипотезу, механизм. Кому и почему результаты исследования могут быть полезны?
2. (5 баллов) Генеральная совокупность, ее географические и временные рамки с обоснованием их выбора. Способ формирования выборки, обоснование репрезентативности, ее размер.
3. (5 баллов) Источник данных (например, если анализ доступных данных, то какие могут быть потенциальные источники и переменные, если анкетирование – то какие вопросы войдут в анкету). Подумайте, какие дополнительные переменные можно использовать для анализа. Влияет ли на ваш исход что-то еще, помимо основной переменной интереса, и каким образом?
4. (5 баллов) Типы переменных, типы используемых графиков. Какие показатели и модели будут посчитаны/оценены для ответа на главный исследовательский вопрос?
5. (5 баллов) С какими потенциальными трудностями вы можете столкнуться в процессе проведения исследования?

Решение и критерии оценки

Участник получает полный балл (5 баллов) за пункт, если каждый компонент пункта построен без ошибок. Участник получает 0 баллов за пункт, если в нем пропущены компоненты и/или компоненты построены неверно, с логическими ошибками. Участник получает 0 баллов за задание, если решение написано сплошным текстом (пункты не пронумерованы).