Домашнее задание #2

Тема: Анализ данных и проверка статистических гипотез

Вопрос 1: В чем различие между зависимыми и независимыми выборками?

Когда мы работаем с одной выборкой, мы знаем, что нам нужно выбрать случайную выборку из популяции, измерить статистику этой выборки, а затем выдвинуть гипотезу о популяции на основе этой выборки. Когда мы работаем с двумя независимыми выборками, мы предполагаем, что если выборки будут выбраны случайным образом (в случае медицинского исследования, субъекты будут случайным образом отнесены к группе), то эти две выборки будут меняться только случайно, и разница не будет статистически значимой. Короче говоря, когда мы имеем независимые выборки, мы предполагаем, что данные из одной выборки не влияют на другую.

Независимые выборки могут иметь место в двух сценариях:

- При тестировании разницы средств между двумя фиксированными популяциями мы проверяем различия между выборками из каждой популяции. При случайном отборе обеих выборок мы можем сделать выводы о популяциях
- При работе с субъектами (людьми, домашними животными и тд), если мы выбираем случайную выборку, а затем случайным образом присваиваем половину субъектов одной группе, а половину другой, мы можем сделать выводы о популяциях

Зависимые выборки немного отличаются. Две выборки данных являются зависимыми, когда каждый score в одной выборке сопряжена с определенным score в другой выборке. Эти типы выборок связаны друг с другом.

Зависимые выборки могут возникнуть в двух ситуациях:

- В одном случае группа может быть измерена дважды, например, в ситуации до и после теста (баллы за тест до и после урока)
- В другом сценарии наблюдение в одной выборке совпадает с наблюдением во второй выборке

Вопрос 2: Когда применяются параметрические статистические критерии, а когда - их непараметрические аналоги?

Причины использования параметрических тестов

Причина 1: Параметрические тесты могут хорошо работать с перекошенными и ненормальными распределениями

Параметрические тесты могут хорошо работать с непрерывными данными, которые не являются нормальными, если вы удовлетворяете требованиям к размеру выборки, приведенным в таблице ниже.

Параметрический анализ	**Руководство по размеру образца для нестандартных данных**
1-образец t	Более 20
2-образец t	Каждая группа должна быть больше 15
Односторонняя ANOVA	● 2-9 групп, то каждая группа должна быть больше 15. ● 10-12 групп, то каждая группа должна быть больше 20.

Причина 2: Параметрические тесты могут хорошо работать, когда распространение каждой группы различно

Хотя непараметрические тесты не предполагают, что ваши данные следуют нормальному распределению, у них есть и другие предположения, которые могут быть трудновыполнимы. Для непараметрических тестов, которые сравнивают группы, общим предположением является то, что данные для всех групп должны иметь одинаковый разброс (дисперсию). Если ваши группы имеют различный разброс, непараметрические тесты могут не дать достоверных результатов.

С другой стороны, если используется тест 2-х проб t или одностороннее ANOVA, можно просто перейти к поддиалогу Options и снять флажок Assume equal variances.

Причина 3: Статистическая сила

Параметрические тесты обычно имеют большую статистическую силу, чем непараметрические тесты. Таким образом, с большей вероятностью обнаружиться значительный эффект, когда он действительно существует.

Причины использования непараметрических тестов

Причина 1: Область исследования лучше представлена медианой

Например, центр искаженного распределения, как и доход, может быть лучше измерен по медиане, где 50% выше медианы и 50% ниже. Если добавить к выборке несколько миллиардеров, то среднее математическое значение значительно увеличится, даже если доход для типичного человека не изменится.

Когда распределение достаточно искажено, на среднее сильно влияют изменения, происходящие далеко в хвосте распределения, в то время как медиана продолжает более точно отражать центр распределения. Для этих двух распределений случайная выборка в 100 из каждого распределения дает значения, которые существенно отличаются, но медианы не сильно отличаются.

Причина 2: У вас очень маленький размер sample

Выборка не соответствует рекомендациям по размеру выборки для параметрических тестов и нет уверенности в том, что есть нормально распределенные данные, следует использовать непараметрический тест. Если действительно маленькая выборка, то можно даже не удостовериться в том, что ваши данные распределены, так как тесты распределения не обладают достаточной мощностью для получения значимых результатов.

Причина 3: Есть данные по порядку, ранжированные данные или отклонения, которые не получается удалить

Типичные параметрические тесты могут оценивать только непрерывные данные, а на результаты могут существенно влиять отклонения. И наоборот, некоторые непараметрические тесты могут обрабатывать обычные данные, ранжированные данные и не могут быть серьезно затронуты отклонениями.