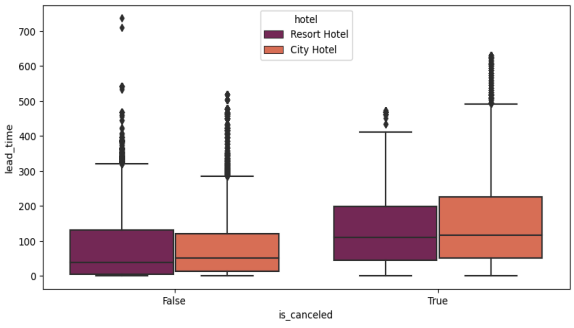
Статистический анализ датасета “Hotel booking demand”

1. Описание датасета

Данный датасет содержит информацию о бронировании номеров в двух португальских отелях в период с 1 июля 2015 до 31 августа 2017 года. Указываются различные данные о бронирующих: дата приезда, количество персон, статус бронирования (отменено или нет), продолжительность пребывания, ADR (Average Daily Rate -- средняя цена ночи в номере). Мы сосредоточим внимание на количестве отмен бронирования и их связи с другими параметрами, представленными в датасете. Здесь и далее данные для различных отелей рассматриваются отдельно, вследствие чего требуется поправка на множественное тестирование Холма-Бонферрони. За уровень значимости принимается α=0.05.

2. Связь времени до заезда с отменой бронирования

В датасете есть информация касаемо времени от бронирования до планируемого заезда (“lead time”). Был проведен t-тест на равенство средних, т.к. выборка имеет большой размер (несколько десятков тысяч записей для каждого отеля), и, согласно центральной предельной теореме, распределение выборочных средних стремится к нормальному.



H_0: среднее время до заезда не отличается в случае отмененной брони

H_1: среднее время до заезда больше в случае отмененного бронирования (односторонняя гипотеза -- разница показана на графике слева)

Результаты тестов представлены ниже. Они позволяют отклонить на уровне значимости α=0.05 гипотезу о равенстве средних для обоих отелей.

	Resort Hotel	City Hotel
t-статистика	-47.18	-91.59
p-value	0.0	0.0

Поправка на множественное тестирование в данном случае не имеет смысла, т.к. значения p-value очень маленькие. Таким образом, люди, отменяющие бронь, в среднем раньше резервируют номера на уровне значимости α=0.05. Также построены доверительные интервалы для среднего времени до заезда (заезд/отмена):

CI = [71.00; 84.67], [122.48; 134.89] (Resort Hotel) ***CI*** = [75.07; 86.34], [142.50; 158.07] (City Hotel)

3. Связь наличия депозита с отменой бронирования

В рассматриваемых отелях присутствует система депозитов -- внесение частичной (‘Refundable’) или полной (‘Non refund’) предоплаты, либо же ее отсутствие (‘No deposit’). Мы решили узнать, отличаются ли количества отмен брони у клиентов с различными типами депозита. **H_0**: наличие депозита и факт отмены бронирования -- независимые переменные. **H_1**: существует зависимость между наличием депозита и фактом отмены бронирования.

Данные по наличию депозита и отмене бронирования для Resort Hotel

Observed	Not canceled	Canceled	Total
No deposit	28749	9450	38199
Non refund	69	1650	1719
Refundable	120	22	142
Total	28818	11100	**39918**

χ^2=4160, p-value=0.0

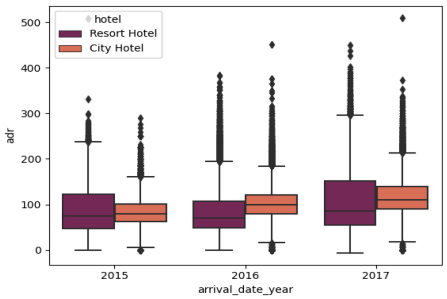
Данные по наличию депозита и отмене бронирования для City Hotel

Observed	Not canceled	Canceled	Total
No deposit	46198	20244	66442
Non refund	24	12843	12867
Refundable	4	14	18
Total	46222	33087	**79309**

χ^2=21324, p-value=0.0

При проведении χ^2-теста было решено не учитывать категорию Refundable deposit по причине маленьких значений у городского отеля (<5). Поскольку p-value << 1, было решено не делать поправку на множественное тестирование, поскольку она не повлияла бы на результаты. В обоих случаях гипотеза о независимости отвергается на уровне значимости α=0.05. Таким образом, факт отмены бронирования связан с наличием депозита. Можно заметить необычайно низкое количество подтвержденных бронирований у клиентов с депозитом. Одно из возможных объяснений -- подобная бронь используется для получения визы на въезд в Португалию[1].

4. Построение доверительного для среднесуточной цены номера



Построим доверительные интервалы (CI) для ADR двух отелей за 2016 год (ADR за три года показаны слева), поскольку это единственный год, данные по которому представлены за все месяцы. Была выбрана выборка размера 100. CI строится следующим образом:

$CI = X \pm z_{a/2} * \frac{s}{\sqrt{n}}$, в данном случае для двух отелей :

$CI_{Resort\ Hotel} = [91.90;\ 98.85]$ $CI_{City\ Hotel} = [102.51;\ 106.59]$, α=0.05

Интерпретация: для большего количества аналогичных случайных выборок полученный доверительный интервал будет включать истинное среднее значение adr за указанный период в 95% случаев.

5. Выводы

- Время от бронирования до заезда у людей, отменивших бронь, значимо отличается от такового у подтвердивших бронирование.
- Отмена бронирования связана с наличием депозита при бронировании
- Мы на 95% уверены, что истинные средние для ADR за 2016 год лежат в интервалах $[91.90;\ 98.85]$ для курортного и $[102.51;\ 106.59]$ для городского отелей.

[1] https://doi.org/10.1177/1938965519851466