Introduction to Data Analysis - Assignment 2

Dataset « ontime_february»

The dataset "ontime_february[1]" compares the on-time arrival performance of airlines in Florida in February. We want to evaluate the performance of the airlines and the airports of departure, understand the reasons for delays and cancellations through this dataset.

Import the "ontime_february" dataset in RStudio and answer the following questions:

Part 1: The airlines performance

1. Create a table named " airline" as described below and fill it:

Airline	Number of flights in the dataset for airline x	Number of delayed flights for airline x	Number of canceled flights for airline x

2. Add two columns to the table, one for « cancelled flights percentage per airline » and the other for « delayed flights percentage per airline »

3. Which airline has the highest delayed flights percentage in the available sample? We call this airline X. Which airline has the lowest delayed flights percentage? We call this airline Y

4. We wonder if the airline X is the one with the lowest delayed flights percentage in general. We address this question by comparing X with each of the other airlines, and compute the p-value.

4.a. Add a new column named "p-value" :

- p-value=NA for the airline X

- p-value: p-value for the test statistic with ,

H_0 : delayed flights proportion for airline = delayed flights proportion for airline X

4.b. Can you reject H_0? for which airline? (α=0.05), Add a new column named "H0"

H_0=0 if we have sufficient evidence that H_0 is false, and H_0=1 if we can't reject H_0

4.c. Add a new column named "IC" that contain the 95% confidence intervals

5. Use a drawing to illustrate the p-value for comparing the proportion of cancelled flights for airline X and the proportion of cancelled flights for airline Y

Part 2: The airports performances

1. Create a table named « airport » as described below and fill it:

OriginAirportID	Number of flights per OriginAirportID	Average DepDelayNew per OriginAirportID	Standard deviation of delay per OriginAirportID

[1] http://www.transtats.bts.gov/

N.B. : if DepDelayNew=NA , replace NA by 0

2. What is the airport with the least departure delay on average based on the sample; we call this airport “A”
3. We wonder if the airport with the least average departure delay in the sample is the one with the least average departure delay in general.

3.a) Add a new column named “SE” that contains the standard error of the average difference between the departure delay in airport “A” and each other airport; for airport A row use NA value.

3.b) Add two new columns that contains the test statistic and p-value for the average difference between the departure delay in airport “A” and each other airport

3.c) Does the data set give us strong evidence that the airport A has the least departure delay on average?

Part3: The reasons of delays and cancellation

1. Import the January dataset and calculate the proportion of cancelled flights for each reason for cancellation. We will use these values as the expected proportions of cancelled flights for each reason in February.

2. Using the sample, calculate the proportion of cancelled flights for each reason in February

3. Do these data provide convincing evidence of an inconsistency between the observed (proportion of cancelled flights for each reason in February dataset) and expected counts observed (proportion of cancelled flights for each reason in January dataset)? Use Chi-square test.

4. Calculate the proportion of delayed flights for each reason in January dataset and February dataset (use CarrierDelay, WeatherDelay, NASDelay, LateAircraftDelay and SecurityDelay columns)

5. Do these data provide convincing evidence of an inconsistency between the observed (proportion of delayed flights for each reason in February dataset) and expected counts observed (proportion of delayed flights for each reason in January dataset)? Use Chi-square test

Appendix: dataset description

AirlineID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
FlightNum	Flight Number
OriginAirportID	Origin Airport ID. An identification number assigned by US DOT to identify a unique airport.
DestAirportID	Destination Airport ID. An identification number assigned by US DOT to identify a unique airport.
DepDelayNew	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
ArrDelayNew	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
Cancelled	Cancelled Flight Indicator (1=Yes)
CancellationCode	Specifies The Reason For Cancellation :A:"Carrier", B:"Weather", C:"National Air System", D:"Security"
CarrierDelay	Carrier Delay, in Minutes
WeatherDelay	Weather Delay, in Minutes
NASDelay	National Air System Delay, in Minutes
LateAircraftDelay	Late Aircraft Delay, in Minutes
SecurityDelay	Security Delay, in Minutes