quiz_01

August 26, 2025

1 Quiz 1

---

1. **Data Classification** *(5)*

   Given a flights dataset detailing the attributes for different flights departing from New York City including Date (year, month, and day), Time (actual & scheduled departure/arrival time), Delay (departure/arrival delay), Flight Details (carrier name, flight number, tail number and origin/destination airport), Journey Details (air time, distance traveled, and total flight time (in hours and minutes)). Classify each variable (Date, Time, Delay, Flight Details, and Journet Details) in the dataset as one of the following: Discrete Quantitative, Continuous Quantitative, Qualitative, and Categorical.

2. **Data Summary** *(5)*

   Using the flights dataset filtered out for John F. Kennedy Airport for Delta Airlines flights on Christmas Eve (24th December), summarise measure of location (mean, median, mode) and dispersion (inter-quartile range and standard deviation) for departure as well as arrival delay.

3. **Probability Analysis** *(5)*

   Prove that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4. **Data Sampling** *(5)*

   a. For the following randomly sampled data from the flights dataset, compute bias and standard error for the estimator on arrival delay, given that population mean of arrival delay is 6.895 minutes (parameter value) . *(3)*

   b. Using the Archery analogy discussed in the class, draw a representative target board to comment upon the accuracy and precision of the estimator. *(2)*

5. **Hypothesis Testing** *(5)*

   Test the following hypothesis for Delta Airlines flights from John F. Kennedy Airport on Christmas Eve

   a. departure delay is greater than 4 minutes

   b. arrival delay is less than -5 minutes

   Given, population mean of departure and arrival delay is 4.04 and -5.16 minutes, respectively; and population sd of departure and arrival delay is 13.092 and 19.996 minutes, respectively.

[1]:
```
# Load Packages
library(dplyr)
library(nycflights13)

# Dataset filtered out for John F. Kennedy Airport for Delta Airlines flights␣
 ↪on Christmas Eve (24th December)
data <- flights %>% filter(origin=="JFK", carrier=="DL", month==12, day==24)
str(data)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


tibble [50 × 19] (S3: tbl_df/tbl/data.frame)
 $ year          : int [1:50] 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013
…
 $ month         : int [1:50] 12 12 12 12 12 12 12 12 12 12 …
 $ day           : int [1:50] 24 24 24 24 24 24 24 24 24 24 …
 $ dep_time      : int [1:50] 615 617 651 657 657 738 757 759 808 818 …
 $ sched_dep_time: int [1:50] 615 615 656 700 700 740 800 800 812 810 …
 $ dep_delay     : num [1:50] 0 2 -5 -3 -3 -2 -3 -1 -4 8 …
 $ arr_time      : int [1:50] 844 915 1144 1017 1045 1019 1059 1110 1045 1046
…
 $ sched_arr_time: int [1:50] 849 850 1141 1040 1026 1054 1109 1058 1111 1043
…
 $ arr_delay     : num [1:50] -5 25 3 -23 19 -35 -10 12 -26 3 …
 $ carrier       : chr [1:50] "DL" "DL" "DL" "DL" …
 $ flight        : int [1:50] 479 404 2275 430 486 440 2190 2381 2596 433 …
 $ tailnum       : chr [1:50] "N3762Y" "N3773D" "N3734B" "N712TW" …
 $ origin        : chr [1:50] "JFK" "JFK" "JFK" "JFK" …
 $ dest          : chr [1:50] "ATL" "ATL" "SJU" "SFO" …
 $ air_time      : num [1:50] 116 129 206 350 294 318 162 163 315 126 …
 $ distance      : num [1:50] 760 760 1598 2586 1990 …
 $ hour          : num [1:50] 6 6 6 7 7 7 8 8 8 8 …
 $ minute        : num [1:50] 15 15 56 0 0 40 0 0 12 10 …
```

```
 $ time_hour      : POSIXct[1:50], format: "2013-12-24 06:00:00" "2013-12-24
06:00:00" …
```

---

[2]:
```r
# Probability Mass Table - Departure Delay
v <- sort(unique(data$dep_delay))
f <- numeric(length(v))
for (r in 1:nrow(data)) {
  z <- data$dep_delay[r]
  i <- which(v == z)
  f[i] <- f[i] + 1
}
df_dep <- data.frame(x=v, f=f/sum(f))
View(df_dep)
```

A data.frame: 21 × 2

x <dbl>	f <dbl>
-9	0.02
-6	0.06
-5	0.06
-4	0.08
-3	0.10
-2	0.12
-1	0.08
0	0.12
2	0.04
3	0.02
5	0.02
6	0.02
8	0.08
11	0.02
17	0.04
19	0.02
20	0.02
25	0.02
29	0.02
37	0.02
66	0.02

---

[3]:
```r
# Probability Mass Table - Arrival Delay
v <- sort(unique(data$arr_delay))
f <- numeric(length(v))
for (r in 1:nrow(data)) {
  z <- data$arr_delay[r]
  i <- which(v == z)
  f[i] <- f[i] + 1
```

```
}
df_arr <- data.frame(x=v, f=f/sum(f))
View(df_arr)
```

A data.frame: 41 × 2

x <dbl>	f <dbl>
-35	0.02
-29	0.02
-28	0.02
-27	0.02
-26	0.04
-25	0.04
-24	0.02
-23	0.02
-22	0.02
-21	0.02
-20	0.02
-19	0.02
-18	0.04
-17	0.02
-16	0.02
-15	0.02
-13	0.02
-12	0.02
-11	0.02
-10	0.02
-9	0.04
-8	0.02
-7	0.02
-6	0.02
-5	0.06
-3	0.02
0	0.02
1	0.02
3	0.06
7	0.04
8	0.02
9	0.02
10	0.02
12	0.02
15	0.02
19	0.02
21	0.02
23	0.02
25	0.02
35	0.02
78	0.02

[4]:
```r
# Randomly sampled data from flights dataset
P <- flights$arr_delay
m <- 30
n <- 1000
z <- mean(P, na.rm=TRUE)
Z <- vector("numeric", m)
for (i in 1:m) {
  set.seed(i)
  I <- order(runif(length(P)))[1:n]
  S <- P[I]
  Z[i] <- round(mean(S, na.rm=TRUE), digits=3)
}
data.frame(estimator=Z)
```

A data.frame: 30 × 1

estimator <dbl>
6.298
6.358
5.903
8.699
5.126
4.782
7.001
4.727
8.028
6.912
6.271
9.190
6.405
4.689
5.129
5.763
7.533
6.298
7.129
7.781
7.725
5.174
6.946
9.530
5.490
7.908
4.631
7.881
6.877
7.304