Desafio 09

Reuben Solomon Katz T22.1

Enunciado

Define-se como quantidade de quilômetros disponíveis de assento (available seat kilometers - ask) como a multiplicação da quantidade de passageiros possíveis de carregar com a quantidade de quilômetros percorridos. Esse dado é de grande importância no setor terceário da aeronáutica, justamente pelo fato de podermos associar o custo de funcionamento do avião com o valor retornado por um passageiro.

Nesse contexto, apartir de uma base de dados da ANAC (agência nacional de Aviação Civil), tenta-se entender essa grandeza para a empresa TAM.

Resolução

Como a base é destribuida

A base é destribuida para cada empresa analisada por essa agência, mostrando detalhes de cada voô, como o ano, mês, destino, origem, etc. No nosso problema, pegaremos apenas a coluna que contém o ask para a empresa LATAM. Com esses dados serão feitas as seguintes ações:

- Média
- Moda
- Mediana
- Variância (não viesada)
- Desvio Padrão (não viesada)
- Box plot
- Hisograma
- QQPlot e QQLine
- Assimetria amostral (não viesada)
- Curstose Amostral (não viesada)
- Testes de Normalidade

Cada ponto pode ser utilizado para entender os voôs dessa empresa e uma forma de cobrar passageiros de maneira efetiva e justa.

Ler dataset e encontrar variáveis

Inicialmente lê-se a base de dados que está no formato excel, retirando-se a linha que não possuem dados. Existem várias maneiras de lidar com a faltas de dados, como imputar, interpolar e deletar. Optamos por esse modelo por simplicidade.

Em seguinda, armazenou-se uma variável que possui o dataset sem repetições (ask_sem_repeticoes), e encontrou-se a distribuição normal do dataset (normal_dist_ask). A distribuição normal representa a probabilidade de observar uma certa grandeza com um valor específico - por isso a integral desse valor é 1. Tal é representado gearlmente pela seguinte função de densidade probabilística(fdp):

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)} e^{\frac{-(x-\mu)^2}{2^2}}$$

onde representa o desvio padrão e é a média. Em R, para encontrar essa fdp, utiliza-se a função pnorm. Além disso, encontrou-se a distribuição empírica das funções (distrib_cumulativa_ask), o qual é a fração de medidas os quais são menores que uma determinada variável. Em outras palavras, ela representa:

$$F_n = \frac{numeroDeElementos < t}{numeroDeElementos}$$

Para encontrar a distribuição das funções empíricas, basta utilizar a função do R ecdf.

```
data = read_excel("consulta6-2018.xlsx",range="Y31775:Y37956")

ask <- unlist(na.omit(data))

normal_dist_ask <- pnorm(ask)

ask_sem_repeticoes = unique(ask)

distrib_cumulativa_ask = ecdf(ask_sem_repeticoes)
```

Media

A media é uma forma de resumir informações de um dataset e representa para onde se concentram os dados de uma distribuilçao, ou seja, é um ponto de equilíbrio. Para se encontrar esse número utiliza-se a seguinte equação:

$$\frac{\sum x_i}{n}$$

Em R, a função que descreve esse valor é a função mean.

```
media_ask <- mean(ask)
```

No nosso problema, encontra-se como média o valor de 1.2584871×10^7.

Mediana

A mediana, assim como a média também descreve o dataet por meio de um único ponto. A diferença entre essas duas grandezas é que essa grandeza não é afetada por assimetria drásticas como a média. Para encontrar esse valor basta encontrar o ponto que divide o conjunto em dois partes iguais.

A função em R que encontra a mediana é median.

```
mediana_ask <- median(ask)
```

No nosso problema, encontra-se como valor da mediana 4.641642×10^6.

Moda

A moda, como os dois valores anteriores também é um ponto que descreve o dataset. Para encontrar tal valor, basta procurar a grandeza que aparece com mais frequência no dataset. Em R não há uma função pronta para tal grandeza, portanto, cria-se uma função para tal.

```r
moda <- function(x) {
   e_unico <- unique(x)
   e_unico[which.max(tabulate(match(x, e_unico)))]
}

moda_ask <- moda(ask)
```

O valor da moda no nosso problema é de 0.

Variância

A variância é uma variável aleatória que mede o "quão longe" os valores se encontram do valor esperado. Para encontrar essa grandeza, pode-se fazer a seguinte fórmula:

$$var = \frac{\sum (X_i - \mu)^2}{n-1}$$

onde

$$\mu$$

é o valor esperado. Pode-se também dividir o valor por n invés de n-1. A diferença entre essas duas grandezas é que a variância é dita não enviesada ao se dividir n-1. Diz se que um valor é não enviesado se seu valor esperado é o próprio parâmetro que se quer estimar. Em R, para se encontrar essa grandeza (o valor não enviesado) utiliza-se a função var.

```r
var_ask <- var(ask)
```

A variância no nosso problema é 4.8577557×10^{14}.

Desvio padrão

O desvio padrão, é definido matematica como a raiz da variância. Essa grandeza representa a dispersão em torno do valor esperado.

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{n-1}}$$

Em R, encontra-se esse valor pela função sd.

```r
desv_pad_ask <- sd(ask)
```

O desvio padrão no nosso problema assume o valor de 2.2040317×10^7.

Desvio Mediano Absolutro

O desvio mediano absoluto é a média das distâncias entre cada dado e a média, como mostra as seguintes fórmulas:

$$MAD = \frac{\sum |x_i - \mu|}{n}$$

Onde μ representa a média aritimétrica. Em R, esse valor é encontrado pela função mad.

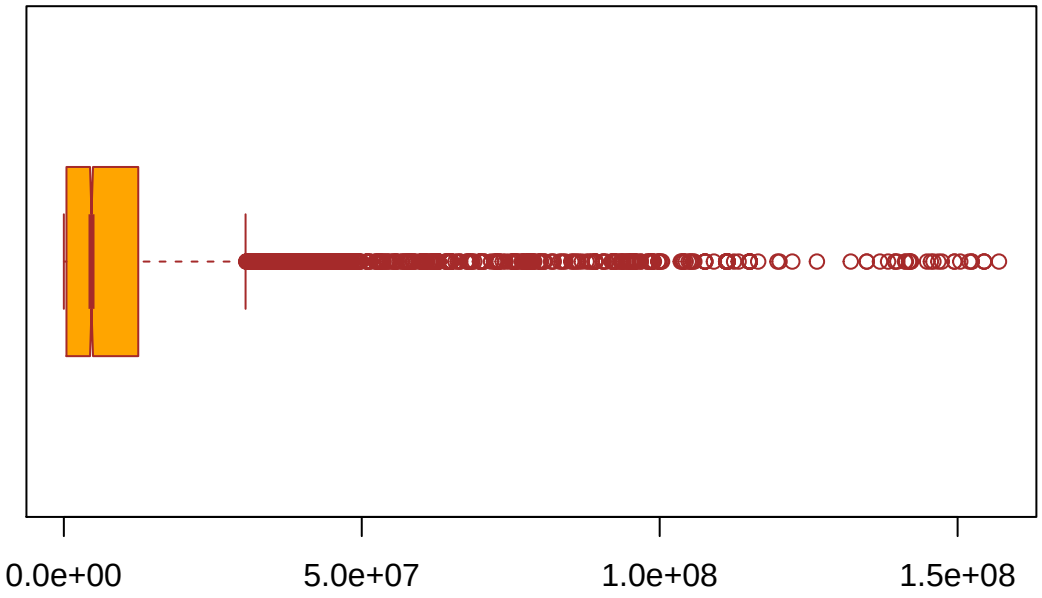
```
mad_ask = mad(ask)
```

No nosso problema, o desvio mediano absoluto é 6.7465209×10^6.

Box Plot de Dados

O box plot é um grafo que indica de maneira sucina como estão distribuidos o meu base de dados. A vantagem desse grafo em relação a hisogramas e plotagens de densidade é que este é menor em tamanho, o que facilita sua comparação com outros grafos "box plot". Em R, plota-se essa distribuição por meio da função boxplot.

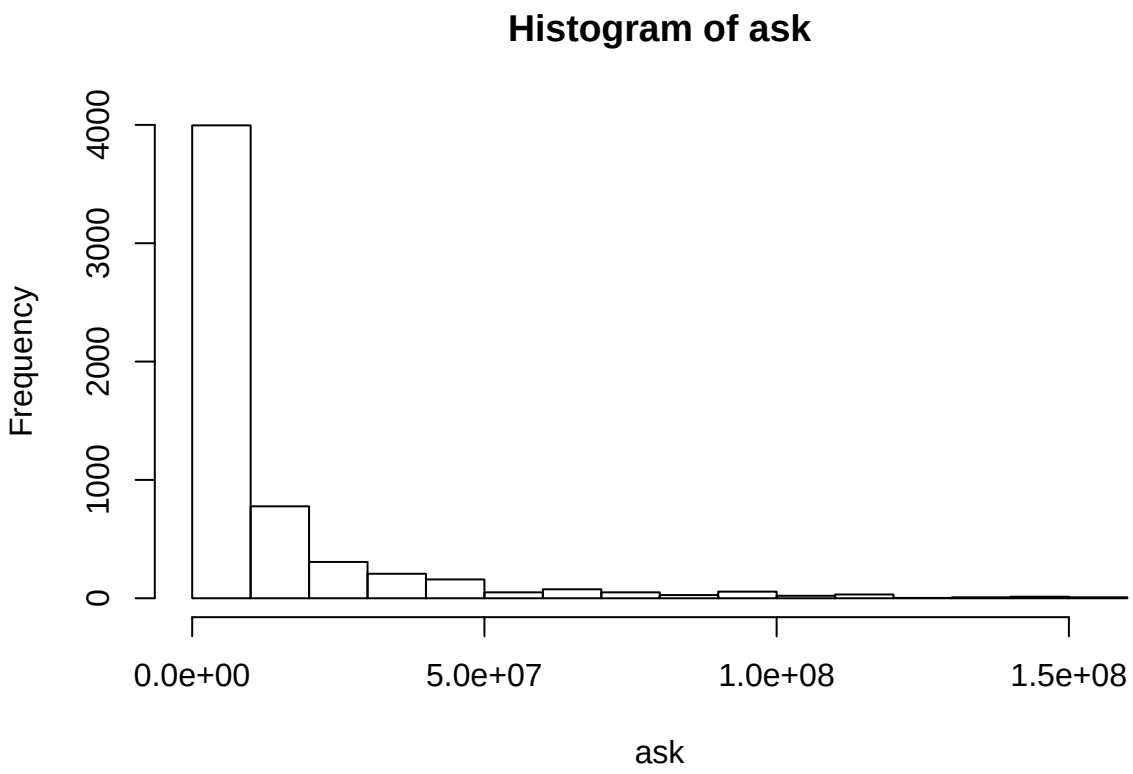
```
boxplot(
  ask,
  col = "orange",
  border = "brown",
  horizontal = TRUE,
  notch = TRUE)
```



Histograma de dados

Um histograma, assim como o box plot, é uma maneira de visualizar a distribuição de dados de um dataset. Cada barra de um histograma representa a frequência em um dado intervalo. Em R, para encontrar essa representação, utilizar-se o comando hist.

```
hist(ask)
```

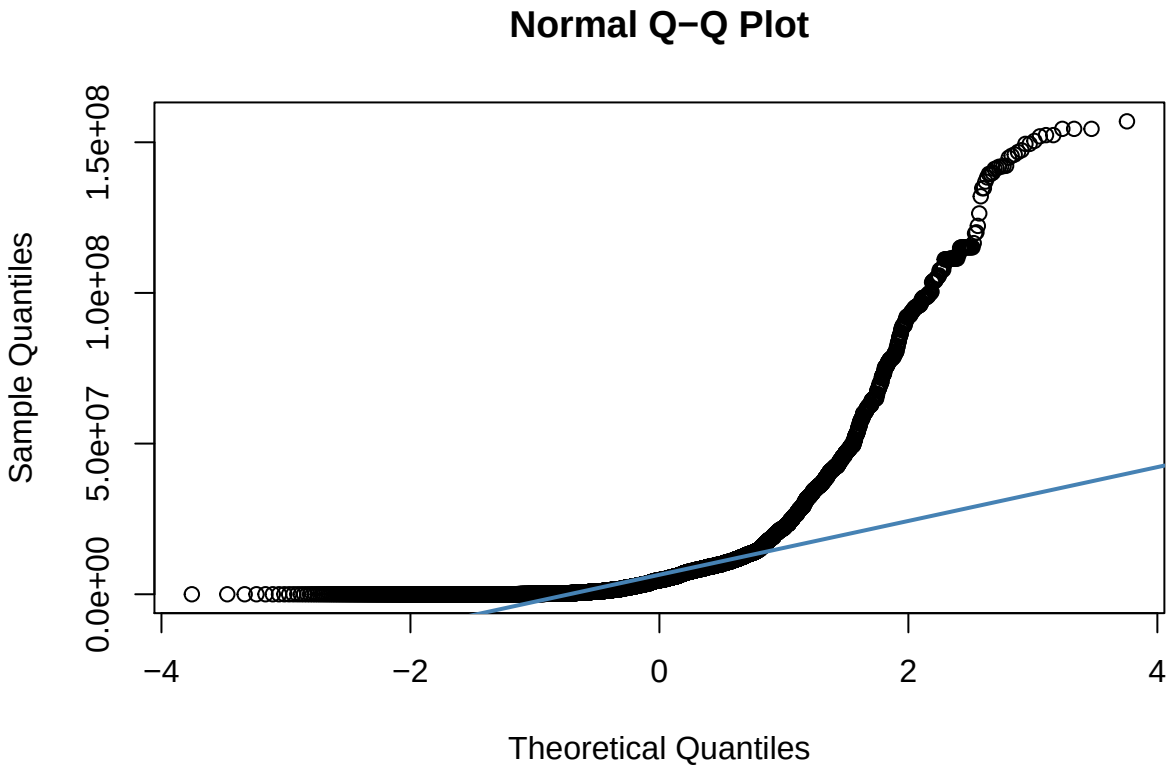


Esse grafo tem utilidade pelo fato de dar uma estimativa onde os valores de um dataset estão concentrados.

QQPlot

O qqplot (quantile-quantile plot) é um método gráfica para ajudar a mostrar se um dado dataset veio de uma distribuição normal ou uma distribuição exponencial. Para fazer isso, compara-se duas funções de densidade por meio de trechos denominados quantiles, que são uma parte de uma dataset em ordem crescente. Se a figura mostrada formar uma reta (ou algo parecido), pode-se esperar que a distribuição é normal, caso contrário, provavelmente a distribuição não é normal. Em R, o comando para plotar o qqplot é qqplot ou qqnorm. A diferença é que o qqnorm escolhe o meu "quantile", enquanto no qqplot eu escolho o quantile utilizado. O qqline é apenas uma referência.

```
{
  qqnorm(ask)
  qqline(ask, col = "steelblue", lwd = 2)
}
```



Assimetria amostral

Uma maneira de avaliar a assimetria de uma amostra é por meio da assimetria(skewness em inglês) e é dado pela seguinte equação:

$$v = \frac{n}{(n-1)(n-2)} \sum (\frac{X_i - \bar{X}}{\sigma})$§$$

onde σ representa o desvio padrão não viesado.

- v>0: Se v>0, então a distribuição tem uma cauda direita (valores acima da média) mais pesada;
- v<0: Se v<0, então a distribuição tem uma cauda esquerda (valores abaixo da média) mais pesada;
- v=0: Distribuição aproximadamente simétrica;

```
asi_ams_ask <- skewness(ask)
```

Em R, a função skewness encontra essa grandeza. No nosso problema, encontra-se um valor positivo de 3.1334548, o que mostra grandezas com mais valores acima da média.

Curtose amostral

A curtose amostral é uma forma da caracterizar o achatamento da curva de uma fdp. Ela é dada pela seguinte equação:

$$v = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum (\frac{X_i - \bar{X}}{\sigma}) - 3\frac{(n-1)š}{(n-2)(n-3)}$$

Similarmente à assimetria amostral, analisa-se a curtose pelo seu sinal:

- v>0:fdp mais alta que a distribuição normal;
- v<0: fdp mais achatada que a distribuição;
- v=0: Mesmo achatamento que a distribuição normal;

```
curtose_ask <- kurtosis(ask)
```

Em R, a função para encontrar essa grandeza é a função kurtosis. No nosso problema, possui um valor positivo de 14.322016, portanto, a sua fdp é mais alta que sua distribuição normal.

Testes de normalidades

Os testes de normalidade são testes usados para determinar se um dataset é bem modelada por uma distribuição normal ou não. No nosso problema, utilizou-se 4 testes:

- Teste de Kolmogorov-Smirnov
  - É uma forma pior que os outros métodos testados
  - No nosso problema este mostra que nosso dataset **não** é bem descrito de uma distribuição normal.
- Teste de Shapiro-wilk
  - No nosso problema este mostra que o nosso dataset é **não** bem descrito por uma distribuição normal.
- Teste de Lillifors
  - No nosso problema, este mostra que nosso dataset **não** é ser adequado.
- Teste de Anderson-Darling
  - No nosso problema, este mostra que nosso dataset **não** é uma distribuição normal.

Para fazer essas análises é preciso analisar duas grandezas:

1. p-value: Denomina a probabilidade a hipótese nula ocorrer. A hipótese nula é uma afirmação de que todas os dados são independentes.
2. d-statistics: É a maior distância em modulo entre duas fdps. Quanto mais perto de zero é esse número, mais parecidas são as curvas.

Em R, utilizou-se as seguintes funções para encontrar-se essas grandezas:

```
ks_ask = ks.test(ask_sem_repeticoes,distrib_cumulativa_ask)

sw_ask = shapiro.test(ask[1:5000])

ad_ask = ad.test(ask)

lilli_ask = lillie.test(ask)
```