Probabilitat i Estadística

23 de desembre del 2016

Grup 11J

José Antonio González Alastrué

Sumari

Sumari de figures

1.Resum

1.1.Objectiu

Estimar per interval la diferència del temps de retard dels vols de Ryanair i Vueling.

1.2.Mètode

Obtenció de les dades a través d'un *script* en R a partir del servidor de la web FlightStats.

1.3.Resultats

No es correcto, no son minutos, es un factor

Amb una confiança del 95%, la diferència del temps de retard es troba entre 1,38 i 1,9 minuts, sent la companyia amb menys retard Vueling ($p = 9{,}117 \cdot 10^{-6}$).

El retraso de uno está entre 1.38 y 1.9 veces más que el otro (la diferencia real está en torno a los 3 minutos de media)

1.4.Discussió

Els resultats obtinguts mostren que Vueling pateix menys temps de retard, i que les premisses de normalitat i d'igualtat de variàncies són clares amb la transformació logarítmica.

2.Objectiu

L'objectiu és saber si la companyia afecta al temps de retard d'un vol, i en cas afirmatiu, si les dues línies aèries de baix cost per excel·lència (Vueling o Ryanair) tenen el mateix temps de retard mitjà.

3.Mètode

A través d'un *script* en R (disponible a l'Annex I: *Script* de recollida de dades en R) hem recollit cada vespre les dades del mateix dia des de les 0h fins les 24h. Aquestes dades estan extretes concretament entre el 7 de novembre del 2016 i el 17 de desembre del 2016 ambdós inclosos. Una vegada obtingudes aquestes dades i davant la gran quantitat d'informació, hem extret 100 mostres aleatòries dels logaritmes de la mitjana de 15 valors per Vueling i 80 mostres aleatòries dels logaritmes de la mitjana de 15 valors per Ryanair. Cal dir però, que cap de les dades en cap de les mostres està repetida. Els *scripts* utilitzats per obtenir la mostra es poden trobar a l'Annex II: Altres *scripts* en R i les dades obtingudes a l'Annex III: Dades recollides.

Així doncs, aquestes dades consisteixen en una mostra aleatòria de les sortides des de l'Aeroport del Prat de les companyies Vueling i Ryanair amb destinacions comunes compreses entre les dades indicades anteriorment. Les destinacions són: Birmingham (Anglaterra), Bolonya (Itàlia), Brussel·les (Bèlgica), Budapest (Hongria), Dublín (República d'Irlanda), Edimburg (Escòcia), Eivissa (Illes Balears), Estocolm (Suècia), Fes (Marroc), Hamburg (Alemanya), Las Palmas de Gran Canaria (Illes Canàries), Liverpool (Anglaterra), Londres (Anglaterra), Manchester (Anglaterra), Marràqueix (Marroc), Menorca (Illes Balears), Palma (Illes Balears), París (França), Roma (Itàlia), Sofia (Turquia), Torí (Itàlia) i Varsòvia (Polònia).

3.1.Prova de significació (bilateral)

$H_0 : \mu_R = \mu_V$

$H_1 : \mu_R \neq \mu_V$

3.2.Premises

Donat que les dades obtingudes en primer terme no seguien una distribució normal, hem realitzat un seguit de transformacions per tal de poder treballar seguint la distribució normal.

1. Normalitat per cada mostra obtinguda de cada grup.
2. Igualtat de variàncies entre el conjunt de mostres dels dos grups.
3. Mostra aleatòria.

3.3.Estadístic i distribució de referència

$$\hat{t} = \frac{\bar{y}_R - \bar{y}_V}{s\sqrt{\frac{1}{n_R} + \frac{1}{n_V}}} \sim t_{n_R+n_V-2}$$

$$s^2 = \frac{(n_R - 1) \cdot {s_R}^2 + (n_V - 1) \cdot {s_V}^2}{(n_R - 1) + (n_V - 1)}$$

3.4.Valor per la regió crítica

Si $|T| > t_{n_R+n_V-2,0.975}$, aleshores es rebutjarà la hipòtesi nul·la.

3.5.Interval de confiança

$$IC(\mu_R - \mu_V, 1-\alpha) = \left[\bar{y}_R - \bar{y}_V \pm t_{n_R+n_V-2,0.975} \cdot s\sqrt{\frac{1}{n_R}+\frac{1}{n_V}}\right]$$

4.Resultats

4.1.Descriptiva

La figura 1 mostra que en mitjana, Ryanair té 3 minuts i mig més de retard que Vueling.

	Global		Ryanair		Vueling	
	$n_R + n_V$	Mitjana	n_R	Mitjana	n_V	Mitjana
Retard (minuts)	180	6,957628	80	9,103321	100	5,611387

Figura 1: Descriptiva global i desglossada per companyia.

Com podem veure a la figura 2, els bigotis superiors tenen una longitud més elevada que els inferiors, i això ens fa posar dubte la premissa de normalitat. És per aquest motiu doncs, que les dades no segueixen una distribució normal, pel qual hem fet la transformació logarítmica. Hem aconseguit normalitzar les dades per tal de poder treballar amb elles, tal i com es pot veure a la figura 3 i 4.

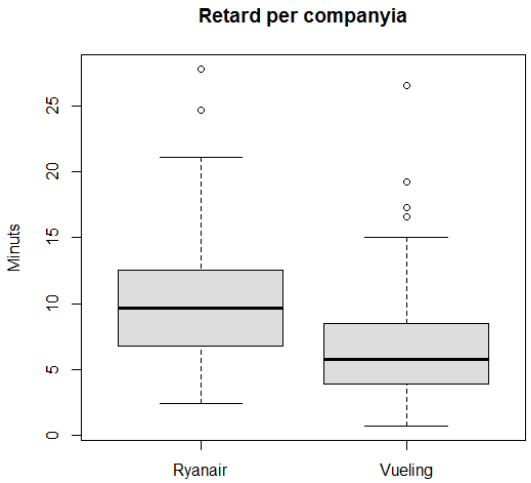

Figura 2: Diagrama de caixes de la mitjana de retards per companyies.

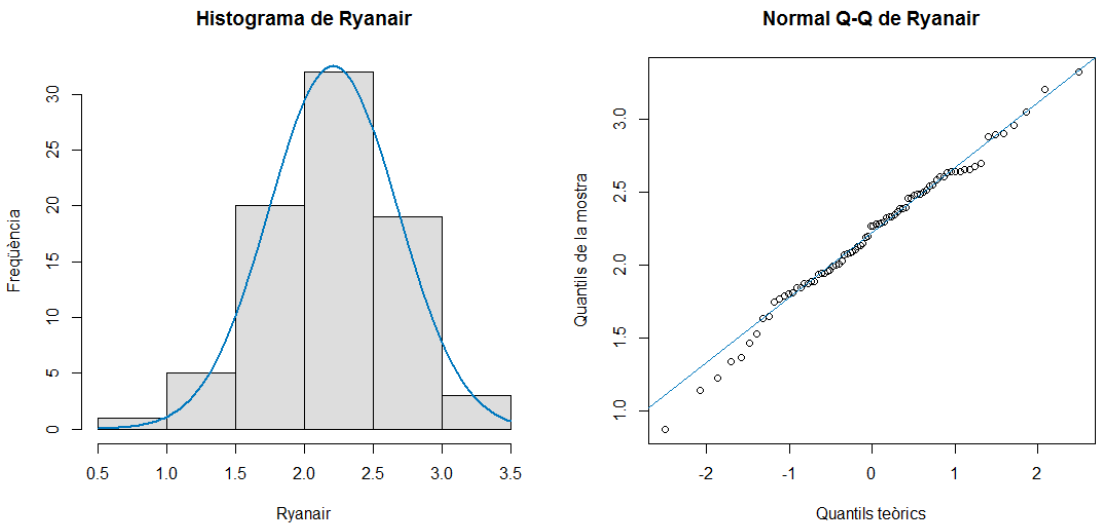

Figura 3: Histograma i normal Q-Q de Ryanair.

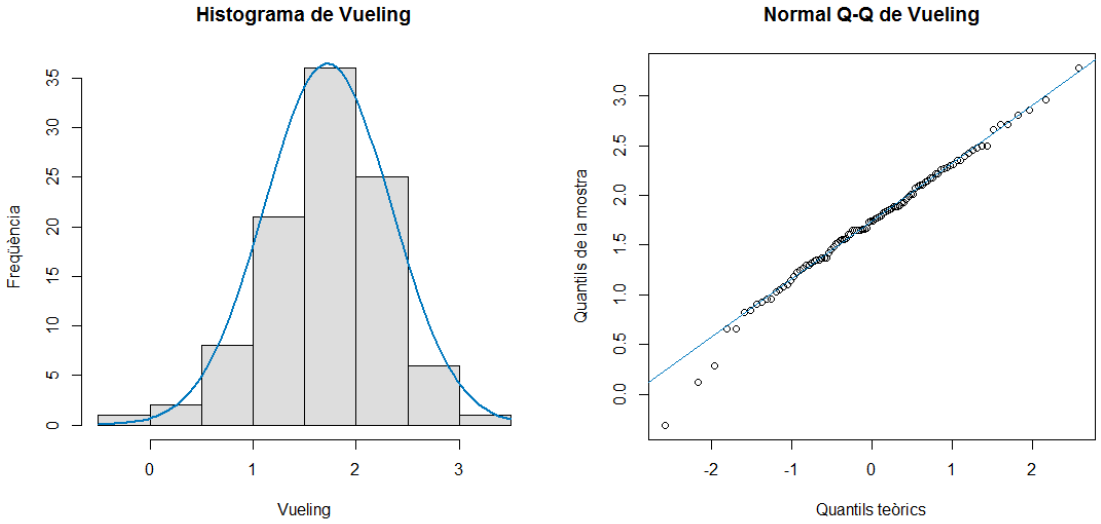

Figura 4: Histograma i normal Q-Q de Vueling.

Veiem però, que les normals Q-Q de cada companyia visibles a les figures 2 i 3 recolzen la premissa de normalitat.

4.2.Prova d'hipòtesi

El resultat de la prova en R és el següent.

```
Welch Two Sample t-test

data: vecR and vecV
t = 6.0171, df = 177.58, p-value = 9.906e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
      0.3251583 0.6425242
sample estimates:
mean of x mean of y
      2.208639  1.724798
```

Faltaba poner "var.equal=TRUE"

El valor de l'estadístic t és 6,0171 i té 178 graus de llibertat ($n_R + n_V - 2$). Amb un nivell de significació del 5%, es rebutja la hipòtesi nul·la d'igualtat de mitjanes de retards entre Ryanair i Vueling ($p = 9{,}906 \cdot 10^{-9}$). L'estimació de la diferència és 0,4838413 ($|\mu_R - \mu_V|$, és a dir, l'esperança de la diferència o la diferència d'esperances per la propietat d'esperança en variables aleatòries) [$IC_{95\%}$ de 0,3252 a 0,6425]. Aquest interval però, és fruit d'una transformació logarítmica, i per tant, hem de desfer la transformació (fent l'exponencial) per poder interpretar-ho correctament [$IC_{95\%}$ de 1,38 a 1,9]. Així doncs, podem afirmar amb un 95% de confiança que Ryanair té un retard mitjà d'entre 1,38 i 1,9 minuts superior al de Vueling.

En realidad, al ser muestras independientes, no es totalmente cierto que la exp(mu_R-mu_V) sea la media del ratio R/V; pero es una aproximación.

5.Discussió

5.1.Conclusió principal

Els resultats obtinguts ens permeten rebutjar amb una forta evidència la hipòtesi d'igualtat (H_0) i per tant podem concloure que la companyia afecta al temps de retard. A més a més, podem assegurar amb una confiança del 95% que Ryanair té una mitjana de retard entre 1,38 i 1,9 minuts superior a Vueling.

5.2.Limitacions

¡Eps!

Les dades obtingudes han estat d'un període determinat i per tant no podem assegurar que no hi hagi una incertesa addicional no contemplada en les mesures estadístiques de l'error aleatori. Per exemple, cal recordar el caos que va provocar Vueling aquest mateix estiu a l'Aeroport del Prat, això vol dir que si haguéssim agafat les dades en aquell període de temps, els resultats no haurien estat tant clars o haurien pogut variar.

5.3.Extrapolació

Cal destacar també, que les premisses de normalitat i igualtat de variàncies són clares amb la transformació logarítmica però no ho acaben de ser sense.

Hubiera estado bien mostrar los datos originales (de todos los dias), y comprobar que efectivamente el muestreo realizado ha acertado (se supone que lo ha hecho)

La interpretación es compleja, ya que se han tomado medias de grupos de observaciones (de varios días, sin relación entre sí). Esto significa que se desea comparar de forma bruta las medias de las dos compañías, evitando el problema de la autocorrelación que presentan los datos temporales.
En general, tratar datos de esta naturaleza presenta muchos problemas complejos, y no es recomendable seguir el ejemplo. Mejor si los datos se obtienen de forma experimental.

Annex I: *Script* de recollida de dades en R

```
library(RCurl)
dest <- c("London", "Brussels", "Rome", "Hamburg", "Milan", "Buda-
pest", "Edinburgh", "Stockholm", "Warsaw", "Sofia", "Bologna", "Palma
Mallorca", "Ibiza", "Menorca", "Turin", "Paris", "Manchester", "Du-
blin", "Liverpool", "Birmingham", "Marrakech", "Fez", "Las Palmas");
for(j in 0:3){
  if(j==0) {
    webpage <- getURL("http://www.barcelona-airport.com/cat/sorti-
des.php?tp=0");
    webpage <- readLines(tc <- textConnection(webpage));
    close(tc);
  }
  else if(j==1){
    webpage <- getURL("http://www.barcelona-airport.com/cat/sorti-
des.php?tp=6");
    webpage <- readLines(tc <- textConnection(webpage));
    close(tc);
  }
  else if(j==2){
    webpage <- getURL("http://www.barcelona-airport.com/cat/sorti-
des.php?tp=12");
    webpage <- readLines(tc <- textConnection(webpage));
    close(tc);
  }
  else{
    webpage <- getURL("http://www.barcelona-airport.com/cat/sorti-
des.php?tp=18");
    webpage <- readLines(tc <- textConnection(webpage));
    close(tc);
  }
  for(i in 0:length(webpage)){
    linia = webpage[i];
    if(!identical(linia, character(0))){
      if(regexpr('Vueling', linia)[1]!=-1){
        liniadest = webpage[i+1];
        pos1 = regexpr(')', liniadest);
        pos2 = regexpr('</td>', liniadest);
        if ((pos1[1]!=-1)&&(pos2[1]!=-1)){
          desti = substr(liniadest, pos1[1]+2, pos2[1]-1);
          if(any(desti %in% dest)){
            liniaweb = webpage[i-1];
            posid1 = regexpr('id=', liniaweb);
            posid2 = regexpr('target', liniaweb);
            posnv1 = regexpr('VY', liniaweb);
            posnv2 = regexpr('</a>', liniaweb);
            if((posid1[1]!=-1)&&(posid2[1]!=-1)&&(posnv1[1]!=-
1)&&(posnv2[1]!=-1)){
              id = substr(liniaweb, posid1[1]+3, posid2[1]-3);
              nvol = substr(liniaweb, posnv1[1]+3, posnv2[1]-1);
              newweb = paste("http://www.flightstats.com/go/FlightSta-
tus/departureDetails.do?id=", id,"&airlineCode=VY&flightNumber=",
nvol, sep="");
```

```
              webpage2 <- getURL(newweb);
              posr = regexpr('statusValue', webpage2);
              status = substr(webpage2, posr[1]+22, posr[1]+41);
              if(substr(status, 0, 1)=="L"){
                status = "0";
              }
              else{
                status = substr(status, 9, regexpr('minutes', sta-
tus)[1]-2);
              }
              linia_hora= webpage[i+2];
              post1 = regexpr('<td>', linia_hora);
              post2 = regexpr('</td>', linia_hora);
              hora = substr(linia_hora, post1[1]+4, post2[1]-1);
              data = paste("Vueling", desti, status, hora, Sys.Date(),
newweb, sep=" | ");
              write(data, file="PE_2016.txt", ncolumns = 1, append =
TRUE, sep="\n");
            }
          }
        }
      }
      else if(regexpr('Ryanair', linia)[1]!=-1){
        liniadest = webpage[i+1];
        pos1 = regexpr(')', liniadest);
        pos2 = regexpr('</td>', liniadest);
        if((pos1[1]!=-1)&&(pos2[1]!=-1)){
          desti = substr(liniadest, pos1[1]+2, pos2[1]-1);
          if(any(desti %in% dest)){
            liniaweb = webpage[i-1];
            posid1 = regexpr('id=', liniaweb);
            posid2 = regexpr('target', liniaweb);
            posnv1 = regexpr('FR', liniaweb);
            posnv2 = regexpr('</a>', liniaweb);
            if ((posid1[1]!=-1)&&(posid2[1]!=-1)&&(posnv1[1]!=-
1)&&(posnv2[1]!=-1)){
              id = substr(liniaweb, posid1[1]+3, posid2[1]-3);
              nvol = substr(liniaweb, posnv1[1]+3, posnv2[1]-1);
              newweb = paste("http://www.flightstats.com/go/FlightSta-
tus/departureDetails.do?id=", id, "&airlineCode=FR&flightNumber=",
nvol, sep="");
              webpage2 <- getURL(newweb);
              posr = regexpr('statusValue', webpage2);
              status = substr(webpage2, posr[1]+22, posr[1]+41);
              if(substr(status, 0, 1)=="L"){
                status = "0";
              }
              else{
                status = substr(status, 9, regexpr('minutes', sta-
tus)[1]-2);
              }
              linia_hora= webpage[i+2];
              post1 = regexpr('<td>', linia_hora);
              post2 = regexpr('</td>', linia_hora);
              hora = substr(linia_hora, post1[1]+4, post2[1]-1);
```

```
                data = paste("Ryanair", desti, status, hora, Sys.Date(),
newweb, sep=" | ");
                write(data, file="PE_2016.txt", ncolumns = 1, append =
TRUE, sep="\n");
              }
            }
          }
        }
      }
    }
}
```

Annex II: Altres *scripts* en R

1.Obtenció de les mostres a partir de les dades

```
library(dplyr)

taula = read.table("PE_2016.txt", header=TRUE, sep="|", quote="");
taulaR = taula[which(taula$Companyia == "Ryanair "), ];

vecR = c();
for(i in 1:80){
    m1 = taulaR[sample(nrow(taulaR), 15), ];
    taulaR = anti_join(taulaR, m1);
    vecR = c(vecR, log(mean((m1$Retard))));
}

taulaV = taula[which(taula$Companyia == "Vueling "), ];

vecV = c();
for(i in 1:100){
    m1 = taulaV[sample(nrow(taulaV), 15), ];
    taulaV = anti_join(taulaV, m1);
    vecV = c(vecV, log(mean((m1$Retard))));
}
```

2.Generació de mitjanes

```
exp(mean(vecR));
exp(mean(vecV));
vecT <- c(vecR, vecV);
exp(mean(vecT));
```

3.Generació de gràfics

3.1.Historiogrames

```
hist(vecR, main="Histograma de Ryanair", xlab="Ryanair", ylab="Fre-
qüència", col="#DDDDDD");
curve(38*dnorm(x, mean(vecR), sd(vecR)), main="Normal Q-Q", add=TRUE,
lwd=2, col="#0078BE");

hist(vecV, main="Histograma de Vueling", xlab="Vueling", ylab="Fre-
qüència", col="#DDDDDD");
curve(56*dnorm(x, mean(vecV), sd(vecV)), add=TRUE, lwd=2,
col="#0078BE");
```

3.2.Normals Q-Q

```
qqnorm(vecR, main="Normal Q-Q de Ryanair", xlab="Quantils teòrics",
ylab="Quantils de la mostra");
qqline(vecR, lwd=1.5, col="#0078BE");
```

```
qqnorm(vecV, main="Normal Q-Q de Vueling", xlab="Quantils teòrics",
ylab="Quantils de la mostra");
qqline(vecV, lwd=1.5, col="#0078BE");
```

3.3.Diagrames de caixa

```
boxplot(exp(vecR), exp(vecV), main="Retard per companyia", ylab="Mi-
nuts", col="#DDDDDD", names=c("Ryanair", "Vueling"));
```

4.Generació de la prova d'hipòtesi

```
t.test(vecR, vecV);
```

```
mean(vecR)-mean(vecV);
```

Annex III: Dades recollides

Vueling				Ryanair		
3,13	5,20	8,20	6,87	13,53	7,00	24,67
5,27	6,27	5,33	11,67	10,27	14,20	7,00
7,93	6,87	3,87	2,33	14,07	7,93	7,33
2,53	5,73	2,60	5,20	12,40	3,93	7,40
5,20	3,93	6,60	8,07	10,67	19,33	14,07
6,40	3,67	8,20	7,07	10,47	6,13	6,07
2,27	4,60	14,33	9,13	5,73	4,33	9,80
3,73	2,60	3,93		11,67	7,47	8,47
12,07	26,53	15,07		12,00	11,67	8,13
1,33	4,27	2,47		6,53	12,07	12,80
3,00	5,27	11,93		10,87	8,07	8,20
5,20	1,93	3,87		6,93	21,13	10,20
12,13	10,93	0,73		10,87	10,27	8,40
4,67	5,73	15,07		6,33	13,27	18,07
4,73	8,80	9,73		8,60	9,87	7,60
1,93	2,80	4,13		12,73	18,27	4,60
6,60	3,47	3,80		9,80	3,13	9,93
17,33	7,33	3,40		5,20	17,80	14,87
4,80	6,20	5,00		7,07	9,67	
10,47	5,67	8,53		27,80	9,00	
3,27	4,73	7,47		14,53	6,33	
3,67	4,40	11,27		14,00	8,00	
7,47	1,13	9,20		9,67	12,20	
2,93	6,47	5,93		11,93	6,60	
8,80	5,87	2,87		6,00	6,60	
5,00	19,27	6,67		13,60	5,13	
6,60	16,60	6,00		11,00	13,93	
5,80	4,53	3,53		3,40	7,13	
5,20	3,93	10,07		5,87	2,40	
8,47	6,33	9,60		6,53	3,80	
9,67	10,53	10,00		14,27	8,93	