Pregunta 1

Finalizado

Puntúa como 6,00

Marcar pregunta

La compañía de taxis "MiTaxi" realizó un estudio para analizar el horario en el que se hacen uso de los servicios de taxi según y el sexo de los usuarios. Con los datos obtenidos a través de una muestra aleatoria de 124 viajes realizados seleccionados al azar durante el último mes, se obtuvo los datos en **viajes.xlsx**.

a) (2.5 puntos) Se sospecha que la duración de los viajes es diferente para hombres y mujeres.

b) (3.5 puntos) Se sospecha que la duración de los viajes varía en los diferentes horarios (madrugada, mañana, tarde, noche).

Pregunta1.R

Rplot02.pdf

a) (0.5 punto) Si los datos no siguen una distribución normal, indique la prueba de hipótesis adecuada para probar si el método es mejor.

Prueba paramétrica U-Mann Whitney con cola a la izquierda para:

H0: Me1=Me2

H1: Me1<Me2

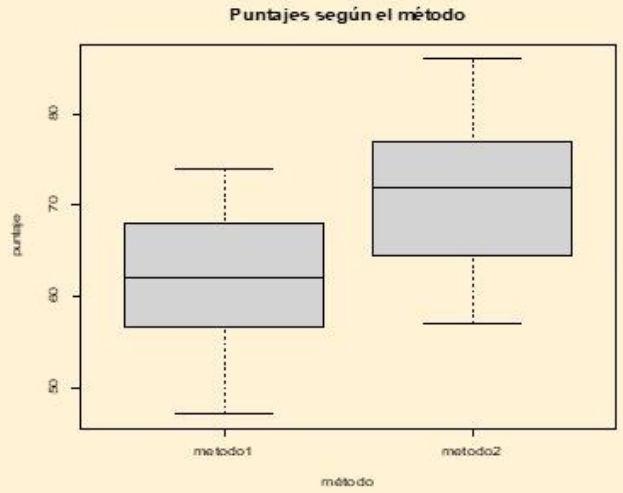
b) (2.5 puntos) Realice la prueba de hipótesis correspondiente y escriba sus conclusiones, con las descripciones del caso.

```
> Tapply(puntaje ~ metodo, median, na.action=na.omit, data=PuntXMet) # medians by group
metodo1 metodo2
     62      72
```

```
> wilcox.test(puntaje ~ metodo, alternative="less", data=PuntXMet)
        Wilcoxon rank sum test with continuity correction
data:  puntaje by metodo
W = 32, p-value = 0.01071
alternative hypothesis: true location shift is less than 0
```

Conclusión: la idiferencia entre ls puntajes es significativa.

```
> Boxplot(puntaje ~ metodo, data=PuntXMet, id=list(method="y"), xlab="método", ylab="puntaje", main="Puntajes según el método")
```



Conclusión: el método 2 está asociado a mejores resultados.

Pregunta 2

Finalizado

Puntúa como 3,00

Marcar pregunta

Se piensa que el método de enseñanza 2 es mejor que el método 1. Para averiguar si es verdad, se seleccionaron 24 estudiantes, de manera aleatoria, se les dividió en 2 grupos a los que se les aplicó los respectivos métodos. Luego, todas estas personas obtuvieron sus puntajes de conocimientos en una prueba oral ante un grupo de evaluadores, obteniéndose la siguiente tabla:

método 1	método 2
47	72
57	62
62	57
52	67
62	72
57	62
56	86
68	74
74	68
62	80
74	86
68	74

a) (0.5 punto) Si los datos no siguen una distribución normal, indique la prueba de hipótesis adecuada para probar si el método es mejor.

b) (2.5 puntos) Realice la prueba de hipótesis correspondiente y escriba sus conclusiones, con las descripciones del caso.

a) (0.5 punto) Indique si las muestras son relacionadas o independientes. Justifique.

Se puede considerar que las muestras son relacionadas, ya que los rios corren paralelo y se encuentran próximos, o sea en la misma región geográfica, bajo los mismos climas a lo largo del año, pues hay medids por mes.

b) (2.0 puntos) Asumiendo normalidad, realice una prueba de hipótesis y escriba las conclusiones que ayuden a decidir si el caudal del rio 1 ofrece menos peligro de desborde que el caudal del rio 2.

Si se asume normalidad, la prueba adecuada es la prueba T para muestras pareadas:
H0: mu1=mu2
H1: mu1>mu2

Antes revisemos las medidas descriptivas:

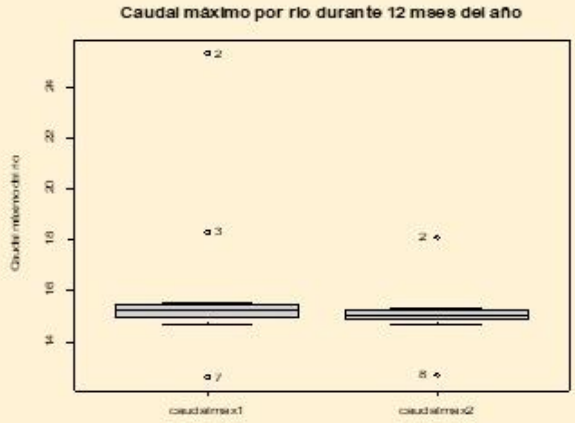
```
> numSummary(caudal[,c("caudalmax1", "caudalmax2", "dif"), drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles", "cv", "skewness"), quantiles=c(0,.25,.5,.75,1), type="2")

              mean       sd       IQR       cv      skewness  0%    25%   50%    75% 100%  n
caudalmax1 16.0583333 3.163846 0.45 0.19702209 2.587879 12.6 14.975 15.25 15.425 25.3 12
caudalmax2 15.1000000 1.173960 0.35 0.07774569 0.882744 12.7 14.875 15.05 15.225 18.1 12

dif        0.9583333 2.329049 0.55 2.43031210 1.922803 -2.1  0.100  0.20  0.650  7.2 12
> with(caudal, (t.test(caudalmax1, caudalmax2, alternative='less', conf.level=.95, paired=TRUE)))
        Paired t-test
data:  caudalmax1 and caudalmax2
t = 1.4254, df = 11, p-value = 0.9091
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
    -Inf 2.165776
sample estimates:
mean difference
      0.9583333
```

c) (0.5 punto) Verifique todo lo que sea necesario para saber si la prueba solicitada fue adecuada.

Boxplot(~ caudalmax1 + caudalmax2, data=caudal, id=list(method="y"), ylab="Caudal máximo del rio", main="Caudal máximo por rio durante 12 mses del año")




```
> normalityTest(~dif, test="shapiro.test", data=caudal)
        Shapiro-Wilk normality test
data:  dif
W = 0.72845, p-value = 0.001607

> normalityTest(~caudalmax1, test="shapiro.test", data=caudal)
        Shapiro-Wilk normality test
data:  caudalmax1
W = 0.63849, p-value = 0.000229

> normalityTest(~caudalmax2, test="shapiro.test", data=caudal)
        Shapiro-Wilk normality test
data:  caudalmax2
W = 0.73273, p-value = 0.001775
```

Conclusión: Hay datos extremos, además no hay normalidad, ni para la diferencia. Por lo que esta no es la prueba adecuada, debería hacerse una prueba no paramétrica.

Mas bien, al contrario: todo parece indicar que debe ser al revés, pues para los rios hay valores extremos por el lado superior, lo que si es indicador de que pueden haber desbordes en ambos rios.

Por otro lado, tanto las medianas como la media son mayores para el rio 1 y su valor extremo es mayor, por lo que parece que el rio 1 tiene más posibilidades de desborde.

Se sugiere realizar una prueba no paramétrica pareada con cola a la derecha.

Pregunta 3

Finalizado

Puntúa como 3,00

Marcar pregunta

En una región geográfica se tienen dos ríos próximos y paralelos. Actualmente, se están tomando medidas para evitar el desborde de los ríos. Para ello, durante los 12 meses del año pasado, se evaluaron los caudales máximos de los ríos.

Los resultados para los dos ríos en los doce meses del año se muestran a continuación:

mes	1	2	3	4	5	6	7	8	9	10	11	12
Caudal máximo del rio 1	15.3	25.3	18.3		15	15.1	15.2	12.6	14.7	14.9	15.4	15.4	15.5
Caudal máximo del rio 2	15.1	18.1	15.1	14.9		15		15	14.7	12.7	14.8	15.2	15.3

a) (0.5 punto) Indique si las muestras son relacionadas o independientes. Justifique.

b) (2.0 puntos) Asumiendo normalidad, realice una prueba de hipótesis y escriba las conclusiones que ayuden a decidir si el caudal del rio 1 ofrece menos peligro de desborde que el caudal del rio 2.

c) (0.5 punto) Verifique todo lo que sea necesario para saber si la prueba solicitada fue adecuada.

a) (2.5 puntos) Se sospecha que la duración de los viajes es diferente para hombres y mujeres.

Sea Y = duración del viaje cuando el pasajeo es hombre
X = duración del viaje cuando el pasajero es mujer

- Verificación de supuesto de muestras independientes: es conforme, pues las muestras de viajes realizados por hombres y mujeres son formadas por viajes diferentes

- **verificación de normalidad de la duración por sexo:**

H0: Y ~ N(muY, sigma2Y) y X ~ N(muX, sigma2X)
H1: Y $\neq$ N(muY, sigma2Y) o X $\neq$ N(muX, sigma2X)

```
> normalityTest(duracion ~ sex, test="shapiro.test", data=v)
  p-values adjusted by the Holm method:
  unadjusted adjusted
F 0.13523    0.27046
M 0.66785    0.66785
```

Conclusión: en ambos casos no se rechaza la normalidad, por lo que se puede aplicar una prueba paramétrica.

- **verificación de homocedasticidad**

H0: sigma2Y/sigma2X = 1
H1: sigma2Y/sigma2X $\neq$ 1

```
> Tapply(duracion ~ sex, var, na.action=na.omit, data=v) # variances by group
      F       M
38.36707 35.36053
```

```
> var.test(duracion ~ sex, alternative='two.sided', conf.level=.95, data=v)
        F test to compare two variances
data:  duracion by sex
F = 1.085, num df = 61, denom df = 61, p-value = 0.751
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6537615 1.8007786
sample estimates:
ratio of variances
          1.085025
data:  duracion by sex
```

Conclusión: el p-valor es mayor que alpha=0.05, además el intervalosde confianza para la razón de varianzas contiene el 1. Por lo que no se rechaza la homocedasticidad.

- **Finalmente, se aplica la prueba paramérica t para la diferencia de medias con varianzas desconocidas, pero iguales (caso 2).**

H0: muY = muX
H1: muY $\neq$ muX

```
> t.test(duracion~sex, alternative='two.sided', conf.level=.95, var.equal=TRUE, data=v)
        Two Sample t-test
data:  duracion by sex
t = -0.65405, df = 122, p-value = 0.5143
alternative hypothesis: true difference in means between group F and group M is not equal to 0
95 percent confidence interval:
 -2.871947  1.445496

sample estimates:
mean in group F mean in group M
      21.06032        21.77355
```

Conclusión: el p-valor es mayor que alpha=0.05, además el intervalo de confianza para la diferencia de medias contiene el 0, por lo que no se rechaza que las medias de la duración de los viajes sean iguales para hombres y mujeres.

b) (2.5 puntos) Se sospecha que la duración de los viajes varía en los diferentes horarios (madrugada, mañana, tarde, noche).

Sea Yi = duración de un viaje realizado en el horario i, i={1=madrugada, 2=mañana, 3=tarde, 4=noche}

- V**erificación de supuesto de muestras independientes:** es conforme, pues las muestras de los viajes son de intervalos del día diferentes, por lo que son independientes
- **verificación de normalidad de la duración por horario:**

H0: para todo i, se cumple: Yi ~ N(muYi, sigma2Yi)
H1: existe un i tal que: Yi $\neq$ N(muYi, sigma2Yi)

```
> normalityTest(duracion ~ horario, test="shapiro.test", data=v)
 p-values adjusted by the Holm method:
          unadjusted adjusted
Madrugada    0.93263        1
Manana       0.91245        1
Tarde        0.96223        1
Noche        0.75809        1
```

Conclusión: en los cuatro horarios no se rechaza la normalidad, por lo que se puede aplicar una prueba paramétrica.

- **Probemos homocedasticidad**
  - H0: sigma2Y1=sigma2Y2=sigma2Y3=sigma2Y4=sigma2Y
    H1: Existe i, tal que: sigma2Yi $\neq$ sigma2Y

```
> Tapply(duracion ~ horario, var, na.action=na.omit, data=v) # variances by group
Madrugada    Manana     Tarde     Noche
 50.08160  15.52868  37.46841  17.23582

> bartlett.test(duracion ~ horario, data=v)
        Bartlett test of homogeneity of variances
```

```
        Bartlett test of homogeneity of variances
data:  duracion by horario
Bartlett's K-squared = 12.411, df = 3, p-value = 0.006099
```

Conclusión: el p-valor es menor que alpha=0.05. Por lo que se rechaza la homocedasticidad.

- **Finalmente, debido a la independencia, normalidad y heterocedasticidad, se aplica la prueba paramérica ANOVA con la variante de Welch.**

H0: muY1=muY2=muY3=muY4=muY
H1: Existe i, tal que: muYi $\neq$ muY

```
> AnovaModel.1 <- aov(duracion ~ horario, data=v)
> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value     Pr(>F)
horario       3    971   323.5   10.96 0.00000206 ***
Residuals   120   3543    29.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> with(v, numSummary(duracion, groups=horario, statistics=c("mean", "sd")))                    mean       sd
data:n
Madrugada   16.94286  7.076836  21
Manana      21.59200  3.940645  20
Tarde       20.25837  6.121144  43
Noche       24.92375  4.151605  40
```

```
> oneway.test(duracion ~ horario, data=v) # Welch test
        One-way analysis of means (not assuming equal variances)
data:  duracion and horario
F = 10.852, num df = 3.000, denom df = 53.109, p-value = 0.00001146
```

Conclusión: el p-valor es menor que alpha=0.05, por lo que se rechaza que las medias de la duración de los viajes sean iguales para los diferentes horarios.

- **por último, analicemos donde está las diferenciaa:**

```
> local({
+   .Pairs <- glht(AnovaModel.1, linfct = mcp(horario = "Tukey"))
+   print(summary(.Pairs)) # pairwise tests
+   print(confint(.Pairs, level=0.95)) # confidence intervals
+   print(cld(.Pairs, level=0.05)) # compact letter display
+   old.oma <- par(oma=c(0, 5, 0, 0))
+   plot(confint(.Pairs))
+   par(old.oma)
+ })
        Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
```

Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = duracion ~ horario, data = v)
Linear Hypotheses:

	Estimate	Std. Err	t value	Pr(>|t|)
Manana - Madrugada == 0	4.649	1.698	2.739	0.0347 *
Tarde - Madrugada == 0	3.316	1.446	2.292	0.1040
Noche - Madrugada == 0	7.981	1.464	5.451	<0.001 ***
Tarde - Manana == 0	-1.334	1.471	-0.907	0.7990
Noche - Manana == 0	3.332	1.488	2.239	0.1169
Noche - Tarde == 0	4.665	1.194	3.909	<0.001 ***

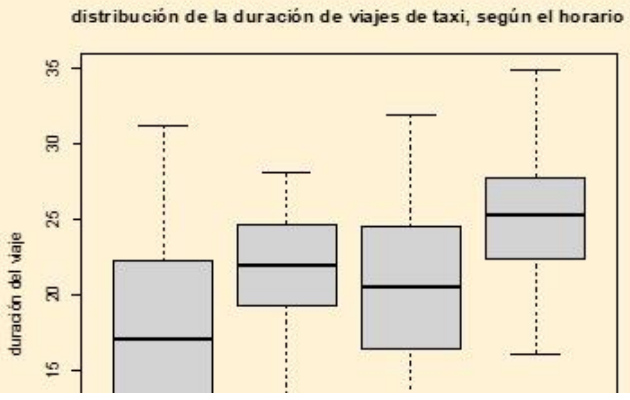
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

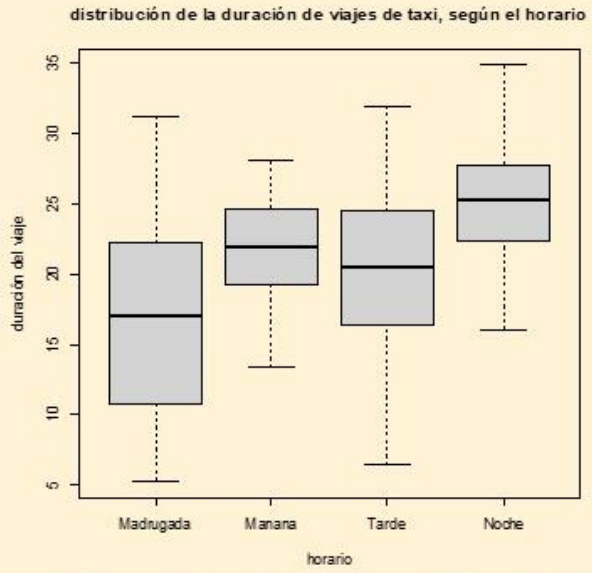
Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = duracion ~ horario, data = v)
Quantile = 2.6009
95% family-wise confidence level

Linear Hypotheses:

	Estimate	lwr	upr
Manana - Madrugada == 0	4.6491	0.2338	9.0645
Tarde - Madrugada == 0	3.3155	-0.4467	7.0777
Noche - Madrugada == 0	7.9809	4.1727	11.7891
Tarde - Manana == 0	-1.3336	-5.1585	2.4912
Noche - Manana == 0	3.3317	-0.5384	7.2019
Noche - Tarde == 0	4.6654	1.5610	7.7697

> Boxplot(duracion ~ horario, data=v, id=list(method="y"), xlab="horario", ylab="duración del viaje", main="distribución de la duración de viajes de taxi, según el horario")





Conclusión: considerando la mediana de la duración, los viajes por la madrugada tienen duración menor, la duración es mayor por la noche.