Tarea Final

Sebastián Andrés Mena Aliaga
Magister en Estadística
Pontificia Universidad Católica de Valparaíso

Abril 2020

1 Pregunta 1

1.1 Parte 1

Se expone el gráfico de boxplot para cada muestra de cada planta, ver figura 1.

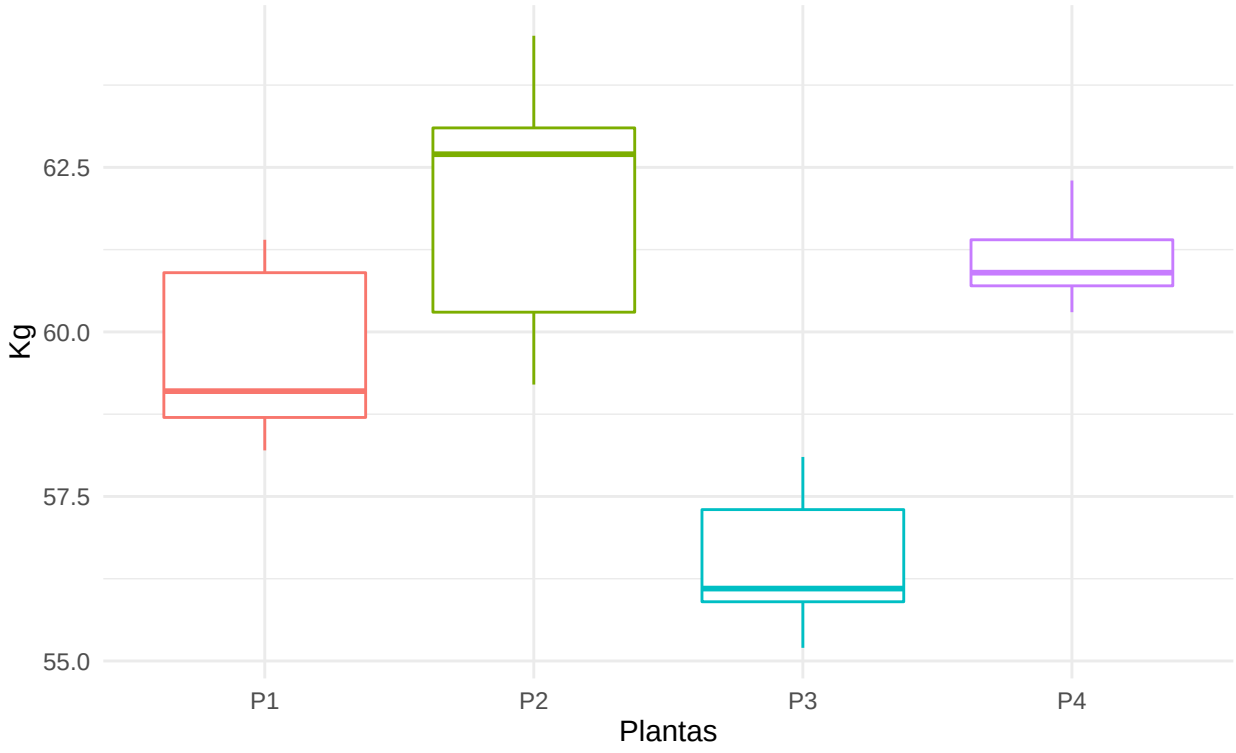

Figure 1: Boxplot de muestras por cada planta

En primer lugar, se realiza un test de Levene para probar la homoceasticidad entre las 4 muestras. La hipótesis nula es que las varianzas poblacionales son iguales, y, al testear con un nivel de significancia $alpha = 0.05$, el p-valor para este caso es 0.397, no hay evidencia para rechazar la hipótesis, por lo tanto, se asume varianzas iguales para los 4 grupos.

Luego, se realiza el test de Kruskal-Wallis para verificar la hipótesis de que, los 4 grupos analizados para este caso, provienen de poblaciones identicas. Es importante señalar que, se asume que las muestras son independientes unas a las otras. A un nivel de significancia $\alpha = 0.05$, el test indica un estadistico $H = 12.855$, $df = 3$ y un $p - valor = 0.00496$, por lo tanto, se rechaza la hipótesis nula. Las muestras no poseen poblaciones identicas.

1.2 Parte 2

Para realizar el test de ANOVA, en primer lugar, se debe evaluar el supuesto de normalidad de cada grupo, para este caso se utiliza el test de Shapiro-Wilk a un nivel de significancia $\alpha = 0.05$ (ver tabla 1). Del test se concluye que no hay evidencia para rechazar la hipótesis de normalidad para ninguna muestra de cada planta.

Además, se debe evaluar el supuesto de que las varianzas son homogéneas entre grupos. Se opta por el test F y se evalua cada par de varianza; cabe destacar que, al validar el supuesto de normalidad anteriormente para cada grupo, es posible realizar el test F. Los resultados (ver tabla 2), a un nivel de significancia $\alpha = 0.05$, indican que no se puede rechazar la hipótesis de varianzas homogeneas entre los grupos.

Planta	W	p-valor
P1	0.891	0.360
P2	0.944	0.694
P3	0.949	0.730
P4	0.948	0.722

Table 1: Test Shapiro-Wilk para cada planta

Planta i	Planta j	F	p-valor
P1	P2	0.425	0.427
P1	P3	1.467	0.720
P1	P4	3.350	0.268
P2	P3	3.453	0.257
P2	P4	7.885	0.070
P3	P4	2.284	0.443

Table 2: Test F para varianzas

2 Pregunta 2

2.1 Parte 1

Para comenzar, se presenta la gráfica de boxplot para las muestras de cada medición, ver figura 2.

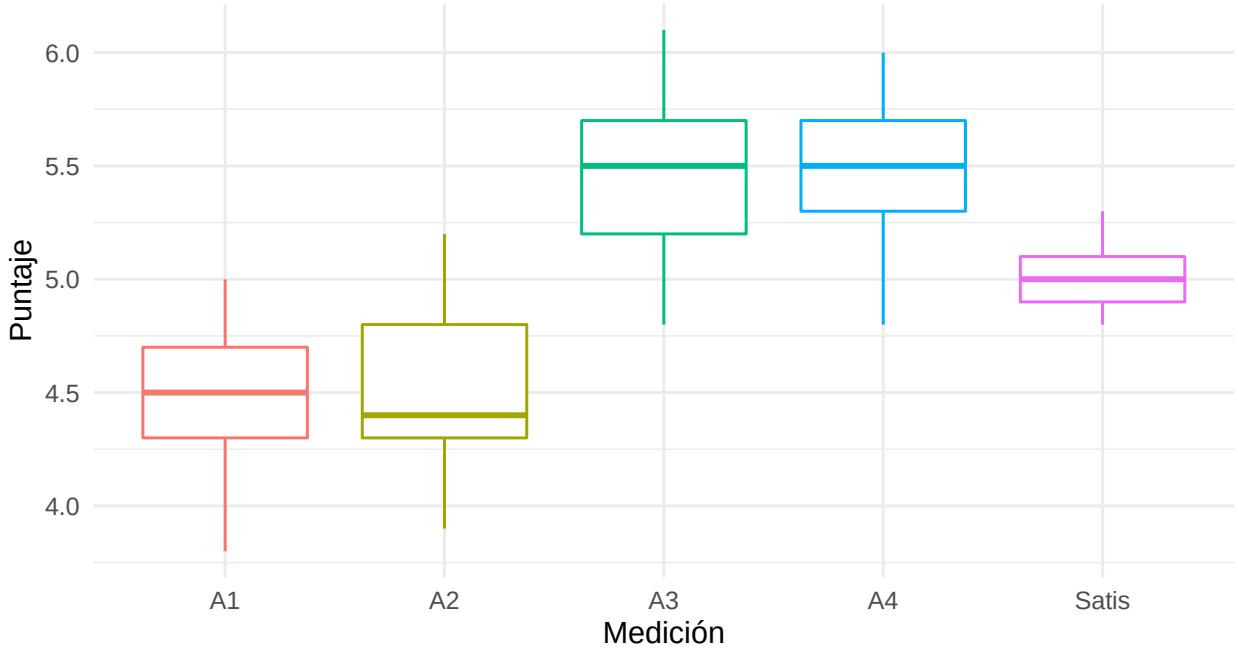

Figure 2: Boxplot de muestras por cada medición: Ambitos y Nivel de Satisfacción General

Se sabe que, para obtener el coeficiente de correlación de Pearson, las muestras a analizar deben proceder de alguna distribución, dado que dicho estadistico es paramétrico. Para ello se realiza el test de Kolmogorov-Smirnov a un nivel de significancia $\alpha = 0.05$, para probar si las mediciones siguen una distribución normal.

Medición	D	p-valor
Satis	0.195	0.004
A1	0.093	0.491
A2	0.154	0.045
A3	0.096	0.449
A4	0.111	0.279

Table 3: Test Kolmogorov-Smirnov para mediciones

Tal como se observa en los resultados de la tabla 3, se decide rechazar la hipótesis nula para las muestras de medición Satis y A2, por lo tanto, no es posible realizar el test de Pearson para evaluar correlación.

Debido que no se validan los supuesto de normalidad para todas las muestras, se opta por utilizar el coeficiente de Spearman no parametrico para las correlaciones. Al realizar el test de correlación a través del método Spearman (ver tabla 4), se rechaza H_0, osea, en favor a que $\rho \neq 0$ para los casos A1, A3 y A4 $\sim$ Satis, además, se puede asegurar a un nivel de significancia $\alpha = 0.05$ que sus correlaciones respectivas son 0.259, 0.372 y 0.382. Para el caso de A2 $\sim$ Satis, no hay evidencia necesaria, para el mismo α, para rechazar que $\rho = 0$, osea, se asume que no hay correlación entre dichas varibales.

Respecto a la pregunta, la variable que presenta mayor correlación respecto a Satis es A4.

Prueba	S	p-valor	ρ
A1 $\sim$Satis	63203	0.020	0.259
A2 $\sim$Satis	73532	0.222	0.138
A3 $\sim$Satis	53619	0.001	0.372
A4 $\sim$Satis	52696	0.001	0.382

Table 4: Test de correlación por método Spearman

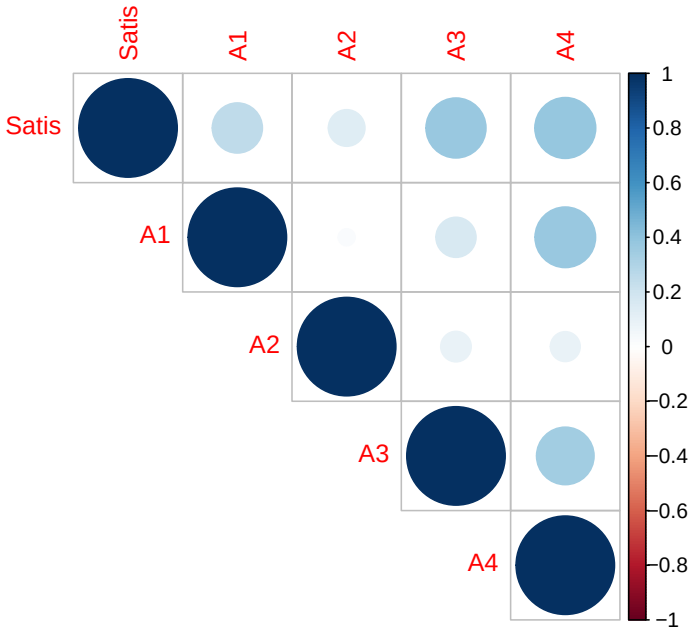

Figure 3: Gráfica de correlación por método Spearman

2.2 Parte 2

En terminos generales, los ámbitos, a excepción de A2 que se asume que no hay correlación ($\rho = 0$), tienen una correlación positiva débil respecto la Satisfacción General (Satis), siendo A4 la que tiene la correlación más alta de estas, seguido de A3 y A1, estas conclusiones se puede observar de mejor forma a través de la figura 3 de correlación.

2.3 Parte 3

Para este caso, las evidencias fueron analizadas en la parte 1.

3 Pregunta 3

3.1 Parte 1

Este test es utilizado para comparar poblaciones cuando sus distribuciones no satisfacen las condiciones suficientes para utilizar test parametricos. Es una alternativa valida al t-test, cuando las muestras no se distribuyen normal o cuando el tamaño de la muestra es muy reducido.

La manera de operar de este test es muy similar al de prueba de signos no parametricos, sin embargo, en vez de asignar signos se asigna la diferencia absoluta entre el valor y la mediana, luego se cálcula el rango a cada valor y se eliminan aquellos que son igual a cero. Posteriormente, se separan los rangos por cuales la diferencia fue positiva y negatica, y luego se suman dichos rangos por separado, el mínimo valor de ambas sumas será el estadístico a probar. Se decide el valor crítico a un nivel significativo dado (tabla Wilcoxon). Por último, si el estadístico es mayor al calor crítico, se rechaza la hipótesis nula.

Ejemplo para 1 muestra: Se quiere estudiar si la mediana del total de ventas realizadas al día por un local es 10 o no. Siendo $H_0 : \widetilde{\mu} = 10$ y $H_1 : \widetilde{\mu} \neq 10$. Los datos y el procedimiento se exponen en la tabla 5, como se puede observar en esta, el estadístico seleccionado es 10, el estadístico crítico a un nivel de significancia $\alpha = 0.05$ es 11. Por lo tanto, no hay evidencia para rechazar que la mediana sea de 10 ventas al día.

Muestra	$D(X - \widetilde{\mu}_0)$	**Rango**	**Rangos +**	**Rangos -**
7	-3	5		
8	-2	3.5		3.5
9	-1	1.5		1.5
10	0	-		
10	0	-		
11	1	1.5	1.5	
12	2	3.5	3.5	
14	4	6.5	6.5	
14	4	6.5	6.5	
16	6	8	8	
18	8	9	9	
20	10	10	10	
		Suma	45	10

Table 5: Ejemplo 1: test Wilcoxon para 1 muestra

Ejemplo para 2 muestra pareadas: Cientificos desarrollan una vacuna contra el COVID-19 y para medir su eficacia la prueban en pacientes (ellos han aceptado realizar la prueba). Los cientificos desean probar si los pacientes han demostrado una diferencia en su condicion previa. La hipótesis a probar es $H_0 : \widetilde{\mu}_A = \widetilde{\mu}_B$ y $H_1 : \widetilde{\mu}_A \neq \widetilde{\mu}_B$, y las muestras pareadas (A y B) y procedimiento se muestra en la tabla 6. El estadístico seleccionado es 14.5, y el estadístico crítico a un nivel de significancia $\alpha = 0.05$ es 11. Dado que el estadístico de prueba es mayor al crítico, por lo tanto, no hay evidencia para rechazar la hiótesis nula, se asume que si hay diferencias entre el estado antes y después de los pacientes.

3.2 Parte 2

Se utiliza el test de Wilcoxon en R para los dos ejemplos anteriores.

Ejmplo 1: Se ejecuta el siguiente código:

```
wilcox.test(ejemplo1, mu = 10, alternative = "two.sided", paired = F, exact = F)
```

El resultado:

```
Wilcoxon signed rank test with continuity correction
data: ejemplo1
V = 45, p-value = 0.08253
alternative hypothesis: true location is not equal to 10
```

La variable *ejemplo1* contiene los mismos valores de Muestra de la tabla 5. El valor p de 0.08523 y, a $\alpha = 0.05$, indica que no hay evidencia suficiente para rechazar que la mediana de la muestra sea igual a 10.

Antes (A)	Después (B)	Diferencia	Rango	Rango +	Rango -
3	8	-5	5.5		5.5
2	17	-15	10		10
4	15	-11	7.5		7.5
8	8	0	-		
6	19	-13	10		10
13	11	2	2.5	2.5	
6	7	-1	1		1
11	9	2	2.5	2.5	
16	11	5	5.5	5.5	
9	20	-11	7.5		7.5
7	4	3	4	4	
			Suma	14.5	41.5

Table 6: Ejemplo 2: test Wilcoxon para 2 muestras pareadas

Ejemplo 2: Código ejecutado:

```
wilcox.test(ejemplo21, ejemplo22, mu =0, alternative = "two.sided", paired = T, exact = F)
```

El resultado:

```
Wilcoxon signed rank test with continuity correction
data: ejemplo2A and ejemplo2B
V = 14.5, p-value = 0.2017
alternative hypothesis: true location shift is not equal to 0
```

Las variables ejemplo2A y ejemplo 2B corresponden a "Antes (A)" y "Después (B)" de la tabla 6, respecticamente. El p-valor es 0.2017, y a un nivel de significancia 0.05, no hay evidencia para rechazar que la diferencia de las medianas sean cero.

4 Pregunta 4

Se seleccionan las variables: capacidad de trabajo bajo presión del postulante (TBT), habilidad lógico matemática (HLM), desempeño en expresión escrita del postulante (EE), nivel de dominio de inglés del postulante (DI) y capacidad de trabajo en equipo (TE).

El grupo N1 cuenta con 284 registros y N2 con 156 registros.

Estadístico	Valor
Min.	41.87
1er Qu.	64.20
Mediana	71.09
Media	72.65
Desv. Est.	11.84
3er. Qu.	81.07
Max.	98.60
NA's	0

Table 7: TBT para N1

Estadístico	Valor
Min.	36.62
1er Qu.	64.23
Mediana	71.93
Media	72.63
Desv. Est.	11.90
3er. Qu.	79.87
Max.	98.91
NA's	3

Table 8: TBT para N2

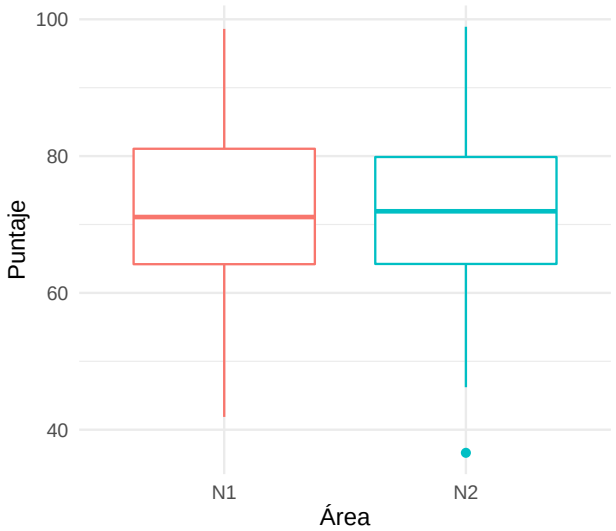

Figure 4: Boxplot de TBT por área

Al analizar el caso de TBT, desde una mirada gráfica (figura 4) pareciera que ambas áreas siguen un comportamiento similar y aparentemente normal, además se observa en el área N3 la presencia de un *outlayer*. Al analizar las tablas 7 y 8, se observa que ambos grupos poseen estadísticos muy similares y, pareciera confirmar a partir de su media y mediana que siguen distribuciones normales. Por otra parte, se observan 3 valores perdidos (NA's) para el grupo N2.

Estadístico	Valor
Min.	41.17
1er Qu.	64.88
Mediana	70.98
Media	72.32
Desv. Est.	11.59
3er. Qu.	80.11
Max.	99.38
NA's	0

Table 9: HLM para N1

Estadístico	Valor
Min.	41.17
1er Qu.	63.97
Mediana	71.06
Media	71.22
Desv. Est.	12.25
3er. Qu.	79.03
Max.	99.83
NA's	2

Table 10: HLM para N2

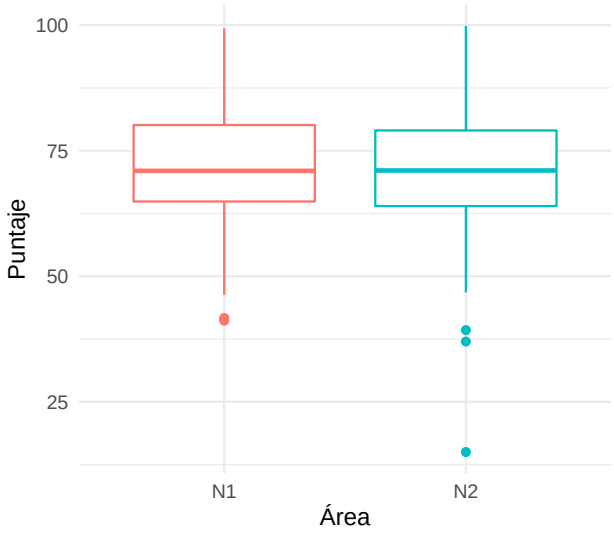

Figure 5: Boxplot de HLM por área

Sobre HLM, la figura 5 podría indicar que las áreas no siguen un comportamiento similar, además de que el área N2 siga una distrubución normal, debido a la presencia de 3 *outlayers*, que podrían generar sesgo. Al analizar las tablas 9 y 10, se observa diferencias levemente más marcadas que el caso anterior que, a diferencia de la mediana y máximo que se aprecian muy similares. Además, se observan 2 valores perdidos (NA's) para el grupo N2.

Estadístico	Valor
Min.	30.05
1er Qu.	55.38
Mediana	63.69
Media	63.68
Desv. Est.	12.00
3er. Qu.	71.91
Max.	93.57
NA's	0

Table 11: EE para N1

Estadístico	Valor
Min.	31.98
1er Qu.	55.68
Mediana	63.98
Media	63.10
Desv. Est.	11.67
3er. Qu.	69.92
Max.	95.20
NA's	2

Table 12: EE para N2

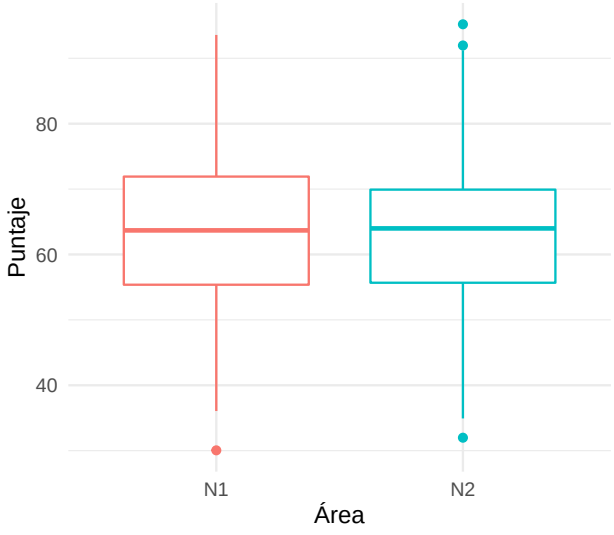

Figure 6: Boxplot de EE por área

Respecto al análisis de EE, se observa en la figura 6 que el grupo N1, debido a su aparente simetria, podría tratarse de una normal, sin embargo, para el grupo N2 su aparente asimetria y su presencia de *outlayers* podrían alejarlo de dicha distribución. Al analizar las tablas 11 y 12, el mínimo, máximo y mediana son muy similares en ambos grupos, y el valor de la media y mediana son muy similares para el grupo N2, además, se observan 2 NA's para el grupo N2.

Estadístico	Valor
Min.	7.78
1er Qu.	31.75
Mediana	45.55
Media	45.48
Desv. Est.	18.45
3er. Qu.	58.64
Max.	90.66
NA's	2

Table 13: DI para N1

Estadístico	Valor
Min.	3.71
1er Qu.	32.38
Mediana	47.21
Media	45.80
Desv. Est.	17.92
3er. Qu.	58.42
Max.	94.38
NA's	1

Table 14: DI para N2

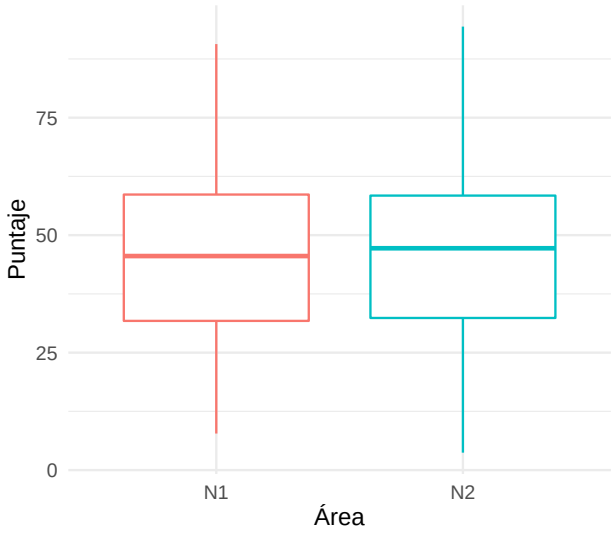

Figure 7: Boxplot de DI por área

Sobre DI, la figura 7 denota que ambos podrían seguir una distribución muy similar y a su vez tratarse ambos de una normal, además, en esta caso no hay presencia de *outlayers*. Al analizar las tablas 13 y 14, se observa similitudes en los estadísticos 1er Qu., mediana, media y 3er Qu., y una leve diferencia en su desviación estandar, además hay similitud medias y medianas respecticas. Se observar 1 NA para el grupo N1 y 2 para el N2.

Estadístico	Valor
Min.	41.34
1er Qu.	60.56
Mediana	66.84
Media	66.69
Desv. Est.	9.37
3er. Qu.	72.70
Max.	92.15
NA's	2

Table 15: TE para N1

Estadístico	Valor
Min.	45.03
1er Qu.	60.89
Mediana	66.64
Media	66.89
Desv. Est.	8.74
3er. Qu.	72.29
Max.	88.64
NA's	2

Table 16: TE para N2

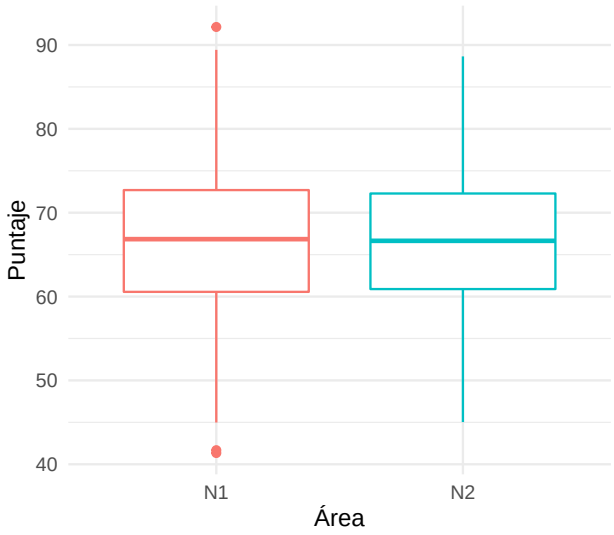

Figure 8: Boxplot de TE por área

Los Boxplots referentes a TE de la figura 7, al igual que el caso anterior, podrían indicar que ambos se distribuyen de manera similar a una normal y, para el caso del grupo N1 se observan 2 *outlayers* a distancias semejantes de la media pero contrapuesta, lo que podría anular su posibilidad de generar asimetría. Al analizar las tablas 15 y 16, se observa similitudes en los estadísticos 1er Qu., mediana, media, 3er Qu. y desviación estandar, además se observa similutud en medias y medianas respectivas. Se observar dos NA's para ambos grupos.

Cabe destacar que, para concluir con mayor propiedad sobre la distribución de los datos, similitud de distribución entre muestras, comaración de medias y varianzas, se debe realizar test de hipótesis apropiados. Lo anterior realizado se conoce como un Análisis Exploratorio de Datos (AED), que tiene como fin conocer de forma previa la naturaleza de estos.

4.1 Parte 2

Tal como se planteó en la parte 1, al realizar un AED se aludió a que las muestras podrían seguir o no una distribución normal, y que podría haber o no igualdad de media y varianza entre áreas al comparar las mismas variables, sin embargo, para fundamentar lo anteriormente dicho con mayor propiedad, es que se recurrer a los test de hipótesis que, entrega una respuesta análitica más fundamentada a una afirmación. A continuación se exponen los resultados y conclusiones. Se asume independencia entre las muestras.

4.1.1 Test de normalidad

Se comienza validando la normalidad de cada unos de las variables escogidas, separado por ambas áreas, a través del test de Kolmogorov-Smirnov. Los resultados expuesto en la tabla 17 y, para un $\alpha = 0.05$, se decide con unanimidad que no hay evidencia para rechazar la normalidad, por lo tanto, se asume dicha distribución para todas las variables por área.

La hipótesis a testear es:

$$H_0 : \text{La muestra se distribuye normal}$$
$$H_1 : \text{La muestra no se distribuye normal}$$

Grupo	Variable	Estadístico	P-Valor
N1	TBP	0.532	0.941
	HLM	0.546	0.928
	EE	0.500	0.965
	DI	0.500	0.965
	TE	0.500	0.965
N2	TBP	0.516	0.954
	HLM	0.513	0.956
	EE	0.532	0.941
	DI	0.516	0.974
	TE	0.519	0.951

Table 17: Test Kolmogorov-Smirnov

4.1.2 Test de homocedasticidad entre variables de N1 y N2

Para evaluar la homogeneidad en la distribución de las varianzas para cada par de variables por área, se decide utilizar el F-test dada su potencia para detectar diferencias muy sutiles entre los valores de la muestra y, además, porque se valida el supuesto de normalidad para dichas muestra. En la tabla 18 se exponen los resultados y, para un $\alpha = 0.05$, se concluye que no hay evidencia para rechazar la hipótesis nula, osea, se supone la igualdad entre sus medias poblacionales.

La hipótesis a testear es:

$$H_0 : \sigma_{N1} - \sigma_{N2} = 0$$
$$H_1 : \sigma_{N1} - \sigma_{N2} \neq 0$$

Variable	Estadístico	Grados Libertad	P-Valor
TBP	0.991	(283, 152)	0.937
HLM	0.895	(283, 153)	0.423
EE	1.057	(283, 153)	0.709
DI	1.060	(281, 154)	0.695
TE	1.150	(281, 153)	0.337

Table 18: F-test

4.1.3 Test de diferencia de medias entre variables de N1 y N2

Se desea evaluar si hay diferencias signigicativa entre las medias poblacionales de las variables por área. Para este caso, al evaluarse anteriormente las condiciones independencia, normalidad y homocedeasticidad, además de contar con un $n > 30$, se decide utilizar el t-test. En la tabla 19 se exponen los resultados y, para un $\alpha = 0.05$, se concluye que no hay evidencia suficiente para rechazar la hipótesis nula, osea, si existe homocedasticidad en cada par de varianza

La hipótesis a testear es:

$$H_0 : \mu_{N1} - \mu_{N2} = 0$$
$$H_1 : \mu_{N1} - \mu_{N2} \neq 0$$

Variable	**Estadístico**	**Grados Libertad**	**P-Valor**
TBP	0.018	435	0.986
HLM	0.925	436	0.355
EE	0.484	436	0.628
DI	-0.176	435	0.860
TE	-0.223	434	0.824

Table 19: t-test

4.1.4 Conclusiones finales

Se puede concluir que no hay diferencias significativas entre los resultados obtenidos para ambas áreas de la empresa. Dicho de otro modo, el desempeño, entre las 5 pruebas seleccionadas, es estadísticamente identica para amas áreas.

5 Referencias

Documentación obtenida para la tarea:

- Kruskal-Wallis test
  https://rpubs.com/Joaquin_AR/219504
- ANOVA análisis de varianza para comparar múltiples medias
  https://www.cienciadedatos.net/documentos/19_anova
- Correlación y Regresión Lineal
  https://rpubs.com/osoramirez/316691
- Análisis de la homogeneidad de varianza (homocedasticidad)
  https://rpubs.com/Joaquin_AR/218466
- Análisis de Normalidad: gráficos y contrastes de hipótesis
  https://rpubs.com/Joaquin_AR/218465
- Prueba de los rangos con signo de Wilcoxon
  https://rpubs.com/Joaquin_AR/218464