Unidad 4 - Actividad 4B Regresión múltiple y logística
Alumno: **Luis Fernando Izquierdo Berdugo**
Materia: **Estadística**
Fecha de Entrega: **10 de Octubre de 2024**

Realiza los siguientes ejercicios del capítulo 4 Fundamentos para la inferencia.

Página	Ejercicios
259	5.11
259	5.15
264	5.27
269	5.39

5.11 Play the piano. Georgianna claims that in a small city renowned for its music school, the average child takes at least 5 years of piano lessons. We have a random sample of 20 children from the city, with a mean of 4.6 years of piano lessons and a standard deviation of 2.2 years.
(a) Evaluate Georgianna's claim using a hypothesis test.

La hipótesis nula H0 es que la media sea mayor o igual a 5 años: $H0 = \mu \geq 5$
La hipótesis alternativa H1 es que la media sea menor a 5 años: $H1 = \mu < 5$

Se usa un nivel de significancia de 0.5

```
> media_muestral <- 4.6
> s <- 2.2
> n <- 20
> mu0 <- 5
> t_estadistico <- (media_muestral - mu0)/(s/sqrt(n))
> t_estadistico
[1] -0.8131156
```

```
> p_value <- pt(t_estadistico, df = n - 1, lower.tail = TRUE)
> p_value
[1] 0.213112
```

Con el valor p mayor a 0.05 no se puede rechazar la hipótesis nula, por lo cual no hay suficiente evidencia para decir que se toman menos de 5 años para aprender piano.

(b) Construct a 95% confidence interval for the number of years students in this city take piano lessons, and interpret it in context of the data.

```
> alpha <- 0.05
> t_critica <- qt(alpha/2, df = n - 1, lower.tail = FALSE)
> intervalo <- c(media_muestral - t_critica * (s / sqrt(n)),
media_muestral + t_critica * (s / sqrt(n)))
> intervalo
[1] 3.570368 5.629632
```

En el intervalo con una confianza del 95%, se puede observar que se incluye el valor de 5 años, apoyando la hipótesis nula.

(c) Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

Si, ambos apoyan la hipótesis nula de que se necesitan 5 años o más para aprender piano.

5.15 Air quality. Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare average air quality between the two years.
(a) Should we use a one-sided or a two-sided test? Explain your reasoning.

Se utilizaria una prueba de dos colas, debido a que se busca comparar la calidad del aire en dos años diferentes, con el objetivo de detectar si ha habido un cambio, ya sea un aumento o una disminución.

La hipótesis nula sería que no hay diferencia en la calidad del aire entre 2013 y 2014, la hipótesis alternativa es que sí hay una diferencia (sin especificar si es un aumento o una disminución).

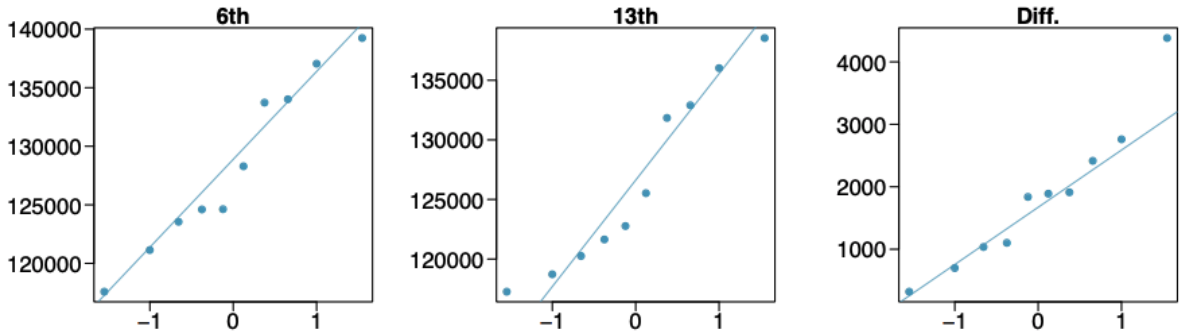
(b) Should we use a paired or non-paired test? Explain your reasoning.

Se usaría una prueba apareada, debido a que las mediciones de la calidad del aire se tomaron en las mismas 25 capitales en ambos años, entonces, al utilizar una prueba apareada, podemos eliminar la variabilidad entre las diferentes capitales y centrarnos en el cambio dentro de cada capital.

(c) Should we use a t-test or a z-test? Explain your reasoning.

Se debe usar una prueba t debido a que no conocemos la distribución de la población de las mediciones de la calidad del aire y el tamaño de la muestra es relativamente pequeño (25 ciudades). Con una prueba t podemos asumir que la distribución de las diferencias entre las mediciones de 2013 y 2014 es aproximadamente normal.

5.27 Friday the 13th, Part I. In the early 1990's, researchers in the UK collected data on traffic flow, number of shoppers, and traffic accident related emergency room admissions on Friday the 13th and the previous Friday, Friday the 6th. The histograms below show the distribution of number of cars passing by a specific intersection on Friday the 6th and Friday the 13th for many such date pairs. Also given are some sample statistics, where the difference is the number of cars on the 6th minus the number of cars on the 13th.



	6^{th}	13^{th}	Diff.
$\bar{x}$	128,385	126,550	1,835
s	7,259	7,664	1,176
n	10	10	10

(a) Are there any underlying structures in these data that should be considered in an analysis? Explain.

Sí, hay una estructura subyacente porque los datos están apareados. Cada observación en el día 6 corresponde a una observación en el día 13 en el mismo lugar y hora.

(b) What are the hypotheses for evaluating whether the number of people out on Friday the 6th is different than the number out on Friday the 13th?

La hipótesis nula H0 dice que no hay diferencia en el número promedio de coches en el viernes 6 y viernes 13. $H0$: $\mu_6 = \mu_{13}$

La hipótesis alternativa H1 dice que si existe una diferencia (aumenta o disminuye) entre el número promedio de coches el día 6 y el día 13. $H1$: $\mu_6 \neq \mu_{13}$

(c) Check conditions to carry out the hypothesis test from part (b).

Independencia: Asumimos que los pares de observaciones (6 y 13) son independientes entre sí. Esto significa que el tráfico en un lugar y día no influye en el tráfico en otro lugar y día.
Normalidad: Los histogramas sugieren que la distribución de las diferencias en el número de coches entre ambos días es aproximadamente normal. Esto es importante porque la prueba t asume que los datos siguen una distribución normal.
Tamaño de muestra: Aunque el tamaño de la muestra es pequeño (n=10), la prueba t es relativamente robusta a la violación de la normalidad, especialmente cuando el tamaño de la muestra no es muy pequeño.

(d) Calculate the test statistic and the p-value.

```
> diferencias <- 128385 - 126550
> t_statistic <- diferencias / (1176 / sqrt(10))
> t_statistic
[1] 4.934336
```

```
> p_value <- 2 * pt(-abs(t_statistic), df = 9)
> p_value
[1] 0.0008085065
```

(e) What is the conclusion of the hypothesis test?

Debido a que el valor p es menor al nivel de significancia estándar de 0.5, se rechaza la hipótesis nula, ya que se tiene evidencia suficiente para decir que hay una diferencia significativa entre el número promedio de coches el viernes 6 y el viernes 13.

(f) Interpret the p-value in this context.

El valor p indica que si la diferencia de tráfico entre ambos días fuera real, habría una probabilidad de 0.08085065% de observar la diferencia en el promedio de coches de ambos días, lo cual es casi imposible.

(g) What type of error might have been made in the conclusion of your test? Explain.

El error podría ser el rechazar la hipótesis nula cuando ésta es cierta, a pesar de que la probabilidad es diminuta, todavía sería una posibilidad.

5.39 Increasing corn yield. A large farm wants to try out a new type of fertilizer to evaluate whether it will improve the farm's corn production. The land is broken into plots that produce an average of 1,215 pounds of corn with a standard deviation of 94 pounds per plot. The owner is interested in detecting any average difference of at least 40 pounds per plot. How many plots of land would be needed for the experiment if the desired power level is 90%? Assume each plot of land gets treated with either the current fertilizer or the new fertilizer.

En este caso se usa la librería "pwr" de R para determinar el número de parcelas que se necesitan para un nivel de poder del 90%.

```
> library(pwr)
> d <- 40 / 94
> sig.level <- 0.05
> power <- 0.90
> sample_size <- pwr.t.test(d = d, sig.level = sig.level, power =
power, type = "two.sample")$n
> sample_size
[1] 117.0232
```

Se necesitan 117 parcelas para obtener un nivel de poder del 90%.