Guía 04 - Ejercicios Evaluados

Alumno: Rodrigo Pereira Yañez

Rut: 16.610.470-k 14/0/2024

Se dice que la vida útil promedio de una batería de teléfono móvil es de 500 ciclos de carga. Para determinar si esto es verdad, se toma una muestra aleatoria de 30 baterías, resultando en una media muestral de 520 ciclos de carga y una desviación estándar muestral de 80 ciclos. Pruebe la hipótesis de que la vida útil de las baterías es diferente a 500 ciclos de carga con un nivel de significancia de α=0.05. Suponga una distribución normal.

Características:

- Prueba para contraste de media, con media poblacional conocida y varianza desconocida.
- Nombre: t-test.
- Tipo de prueba: paramétrica.

Primera pregunta (20 puntos).

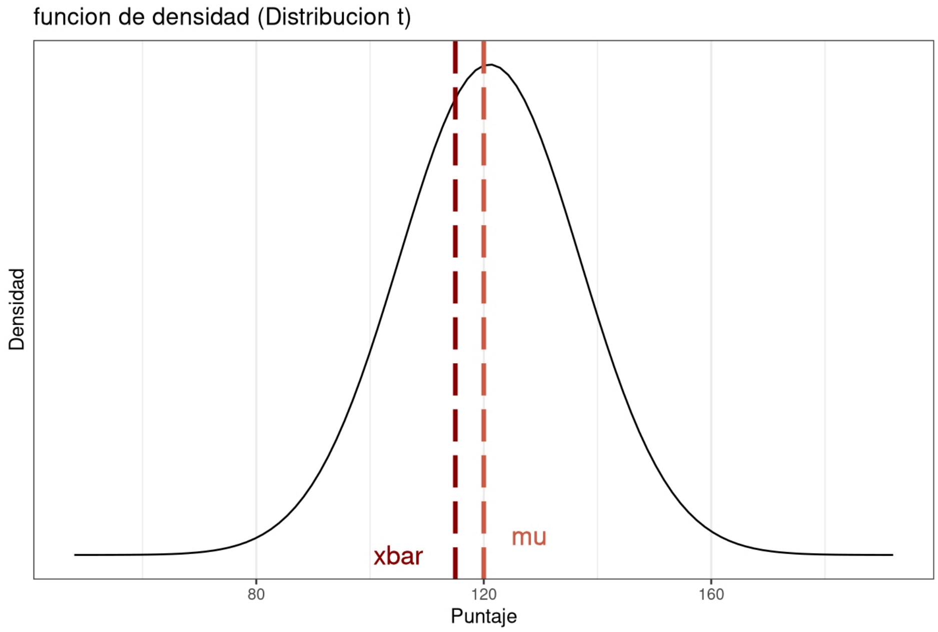
Con lo anterior, busque o cree un ejercicio donde deba aplicar la prueba t, pero para una sola cola.

Respuesta: Un investigador quiere determinar si un nuevo medicamento para reducir la presión arterial tiene un efecto significativo en comparación con un placebo. Según estudios previos, la presión arterial promedio para el grupo de control (placebo) es de 120 mmHg. El investigador selecciona una muestra aleatoria de 25 pacientes que tomaron el nuevo medicamento y obtiene una media muestral de 115 mmHg con una desviación estándar de 10 mmHg.

Pruebe la hipótesis de que la presión arterial promedio de los pacientes que tomaron el nuevo medicamento es menor que la del grupo de control, con un nivel de significancia de α=0.05.

Con los datos proporcionados inicialmente, podemos proceder al cálculo de la distribución de probabilidad asociada al ejercicio.

Gráfico



Código

```r
#================
# 1.- Con lo anterior, busque o cree un ejercicio donde deba aplicar la prueba t,
#     pero para una sola cola. (20 puntos)
#================

library(ggplot2)

grafico_densidad = ggplot(data = data.frame(x = c(48, 192)), aes(x)) +
  stat_function(fun = dt, n = 101, args = list(x = 120, df = 29)) +
  ylab("") + scale_y_continuous(breaks = NULL) + xlab("Puntaje") + ylab("Densidad") +
  ggtitle("funcion de densidad (Distribucion t)") + theme_bw() +
  geom_vline(xintercept = 115, cex = 1.2, colour = "darkred", linetype = "longdash") +
  geom_text(aes(x = 105, label = "xbar", y = 0.0), colour = "darkred", size=5) +
  geom_vline(xintercept = 120, cex = 1.2, colour = "coral3", linetype = "longdash") +
  geom_text(aes(x = 128, label = "mu", y = 0.001), colour = "coral3", size=5)

grafico_densidad
```

Consola

```
> #================
> # 1.- Con lo anterior, busque o cree un ejercicio donde deba aplicar la prueba t,
> #     pero para una sola cola. (20 puntos)
> #================
>
> library(ggplot2)
>
> grafico_densidad = ggplot(data = data.frame(x = c(48, 192)), aes(x)) +
+   stat_function(fun = dt, n = 101, args = list(x = 120, df = 29)) +
+   ylab("") + scale_y_continuous(breaks = NULL) + xlab("Puntaje") + ylab("Densidad") +
+   ggtitle("funcion de densidad (Distribucion t)") + theme_bw() +
+   geom_vline(xintercept = 115, cex = 1.2, colour = "darkred", linetype = "longdash") +
+   geom_text(aes(x = 105, label = "xbar", y = 0.0), colour = "darkred", size=5) +
+   geom_vline(xintercept = 120, cex = 1.2, colour = "coral3", linetype = "longdash") +
+   geom_text(aes(x = 128, label = "mu", y = 0.001), colour = "coral3", size=5)
>
> grafico_densidad
```

Nuestra hipótesis nula, basándonos en los datos señalados en el anunciado, puede ser especificada como:

Hipótesis nula (Ho): μ ≥ 120 mmHg (la presión arterial promedio con el nuevo medicamento no es menor que con el placebo).

Hipótesis alternativa (Ha): μ < 120 mmHg (la presión arterial promedio con el nuevo medicamento es menor que con el placebo).

Ya que es desigualdad, trabajamos con una cola y podemos aplicar la normalización para calcular el valor de t y el intervalo correspondiente.

Código

```r
# Datos del ejercicio sobre vida util
mu = 120     # Valor hipotetico bajo Ho ([mmHg] presion arterial grupo control)
s = 10       # Desviacion estandar muestral ([mmHg] presion arterial)
x_bar = 115  # Media muestral ([mmHg] presion arterial)
n = 25       # Tamaño de la muestra

# Normalizacion
# Calculo del estadistico t
t = (x_bar - mu) / (s / sqrt(n))
t

# Nivel de significancia
alfa = 0.05

# Confianza
confianza = 1-alfa
confianza

# Calculo de intervalo una cola
t_alfa = qt(alfa, df = n-1)
t_alfa
```

Consola

```
> # Datos del ejercicio sobre vida util
> mu = 120     # Valor hipotetico bajo Ho ([mmHg] presion arterial grupo control)
> s = 10       # Desviacion estandar muestral ([mmHg] presion arterial)
> x_bar = 115  # Media muestral ([mmHg] presion arterial)
> n = 25       # Tamaño de la muestra
>
> # Normalizacion
> # Calculo del estadistico t
> t = (x_bar - mu) / (s / sqrt(n))
> t
[1] -2.5
>
> # Nivel de significancia
> alfa = 0.05
>
> # Confianza
> confianza = 1-alfa
> confianza
[1] 0.95
>
> # Calculo de intervalo una cola
> t_alfa = qt(alfa, df = n-1)
> t_alfa
[1] -1.710882
```

El valor de t = -2.5, mientras que el intervalo de t_alpha ≥ -1.710882, en este caso, el t_aplha es solo uno, ya que es una cola.

Lo anterior implica que t está fuera del intervalo de confianza, por lo que podemos rechazar la Ho.

Por lo tanto, podemos decir con un 95% de confianza que la presión arterial promedio con el nuevo medicamento es menor que con el placebo.

Aplicación t-test.

Código

```r
set.seed(100)

# Creacion de muestra artificial
datos = rnorm(n, x_bar, s)

# Aplicacion del test
t = t.test(x = datos, mu = mu, conf.level = confianza)
print(t)
```

Consola

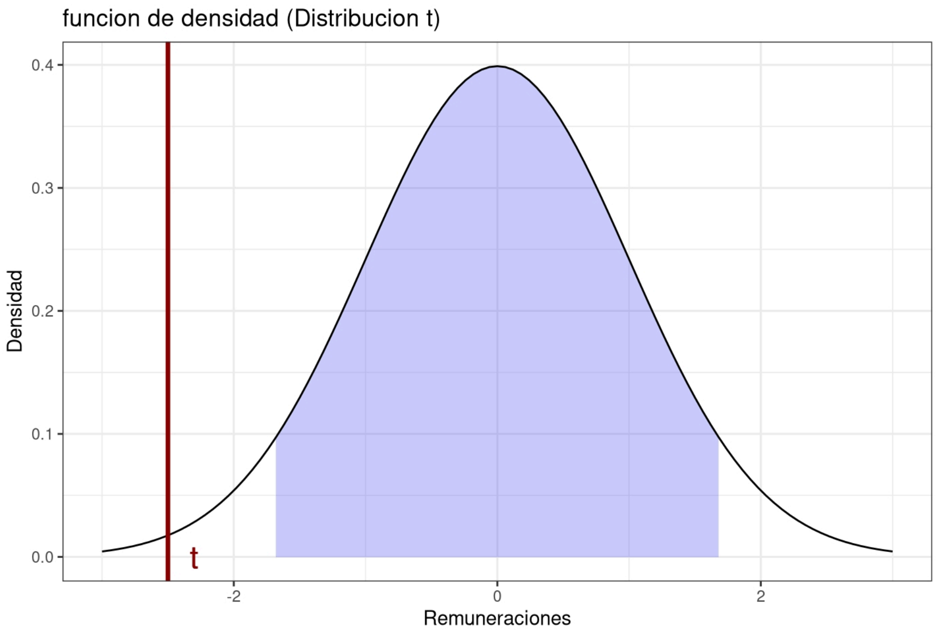
```
> set.seed(100)
>
> # Creacion de muestra artificial
> datos = rnorm(n, x_bar, s)
>
> # Aplicacion del test
> t = t.test(x = datos, mu = mu, conf.level = confianza)
> print(t)

	One Sample t-test

data:  datos
t = -2.7874, df = 24, p-value = 0.01022
alternative hypothesis: true mean is not equal to 120
95 percent confidence interval:
 113.1805 118.9830
sample estimates:
mean of x
 116.0817
```

Lo cual puede ser graficado del siguiente modo.

Gráfico



Código

```
dnorm_limit = function(x) {
  y = dnorm(x)
  y[x < t_alfa | x > -t_alfa] = NA
  return(y)
}

grafico_densidad = ggplot(data = data.frame(x = c(-3, 3)), aes(x = x)) +
  stat_function(fun = dnorm_limit, geom = "area", fill = "blue", alpha = 0.2) + stat_function(fun = dnorm) +
  xlab("Remuneraciones") + ylab("Densidad") + ggtitle("funcion de densidad (Distribucion t)") + theme_bw() +
  geom_vline(xintercept = (x_bar - mu) / (s/sqrt(n)), cex = 1.2, colour = "darkred") +
  geom_text(aes(x = (x_bar - mu) / (s/sqrt(n)) + 0.2, label = "t", y = 0.0), colour = "darkred", size=6)

plot(grafico_densidad)
```

Consola

```
> dnorm_limit = function(x) {
+   y = dnorm(x)
+   y[x < t_alfa | x > -t_alfa] = NA
+   return(y)
+ }
>
> grafico_densidad = ggplot(data = data.frame(x = c(-3, 3)), aes(x = x)) +
+   stat_function(fun = dnorm_limit, geom = "area", fill = "blue", alpha = 0.2) + stat_function(fun = dnorm) +
+   xlab("Remuneraciones") + ylab("Densidad") + ggtitle("funcion de densidad (Distribucion t)") + theme_bw() +
+   geom_vline(xintercept = (x_bar - mu) / (s/sqrt(n)), cex = 1.2, colour = "darkred") +
+   geom_text(aes(x = (x_bar - mu) / (s/sqrt(n)) + 0.2, label = "t", y = 0.0), colour = "darkred", size=6)
>
> plot(grafico_densidad)
```

Segunda pregunta (20 puntos).

Chi-cuadrado fue una de las pruebas no paramétricas vistas en clases, busque y desarrolle un ejercicio donde se utilice chi-cuadrado siguiendo el paso a paso proporcionado al inicio de esta guía.

Respuesta: Una fábrica cuenta con tres máquinas para la producción de un mismo producto. Durante la última semana de producción se han producido 135 artículos. El jefe de producción cree que las máquinas no producen en cantidades similares. Por lo que ha solicitado, clasifiquen cada producto según la máquina que la ha producido.

A continuación se presenta la tabla de frecuencia de las cantidades producidas por cada máquina:

- Máquina A: Producción = 43
- Máquina B: Producción = 53
- Máquina C: Producción = 39

Con un nivel de significancia de $\alpha=0.05$. Pruebe si la cantidad producida es la misma en las 3 máquinas.

Nuestra hipótesis nula, basándonos en los datos señalados en el anunciado, puede ser especificada como:

Hipótesis nula (Ho): La cantidad producida es la misma en las tres máquinas.

Hipótesis alternativa (Ha): La cantidad producida es diferente en las tres máquinas.

Código

```r
#================
# 1.- Chi-cuadrado fue una de las pruebas no paramétricas vistas en clases,
#     busque y desarrolle un ejercicio donde se utilice chi-cuadrado siguiendo
#     el paso a paso proporcionado al inicio de esta guía. (20 puntos)
#================

# Datos del ejercicio
observado = c(43, 53, 39) # Produccion observada
esperado = c(45, 45, 45)  # Produccion esperada

# Normalizacion
# Calculo del estadistico chi-cuadrado
chi_cuadrado = sum((observado - esperado)^2 / esperado)
chi_cuadrado

# Nivel de significancia
alfa = 0.05

```

Consola

```
R 4.3.3 · ~/
> #================
> # 1.- Chi-cuadrado fue una de las pruebas no paramétricas vistas en clases,
> #     busque y desarrolle un ejercicio donde se utilice chi-cuadrado siguiendo
> #     el paso a paso proporcionado al inicio de esta guía. (20 puntos)
> #================
>
> # Datos del ejercicio
> observado = c(43, 53, 39) # Produccion observada
> esperado = c(45, 45, 45)  # Produccion esperada
>
> # Normalizacion
> # Calculo del estadistico chi-cuadrado
> chi_cuadrado = sum((observado - esperado)^2 / esperado)
> chi_cuadrado
[1] 2.311111
>
> # Nivel de significancia
> alfa = 0.05
>
```

Con un total de 135 productos y 3 máquinas, la frecuencia esperada para cada máquina sería: 45. El estadístico chi-cuadrado tiene un valor de: 2.311111

El valor crítico del estadístico chi-cuadrado con 2 grados de libertad y α=0.05 es aproximadamente 5.991465

Como chi-cuadrado < valor crítico del estadístico chi-cuadrado, por lo que se acepta la Ho.

Por lo tanto, podemos decir con un 95% de confianza que as máquinas producen en cantidades similares.

Aplicación qchisq.

Código

```r
# Grados de libertad
df = length(observado) - 1

# Aplicacion del test: Valor critico
valor_critico = qchisq(1 - alfa, df)
valor_critico

# Conclusion de la prueba
result = ifelse(chi_cuadrado > valor_critico,
                "Rechazamos H0: Las máquinas no producen en cantidades similares.",
                "No podemos rechazar H0: Las máquinas producen en cantidades similares.")
print(result)
```

Consola

```
>
> # Grados de libertad
> df = length(observado) - 1
>
> # Aplicacion del test: Valor critico
> valor_critico = qchisq(1 - alfa, df)
> valor_critico
[1] 5.991465
>
> # Conclusion de la prueba
> result = ifelse(chi_cuadrado > valor_critico,
+                 "Rechazamos H0: Las máquinas no producen en cantidades similares.",
+                 "No podemos rechazar H0: Las máquinas producen en cantidades similares.")
> print(result)
[1] "No podemos rechazar H0: Las máquinas producen en cantidades similares."
>
```

DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Tercera pregunta (20 puntos).

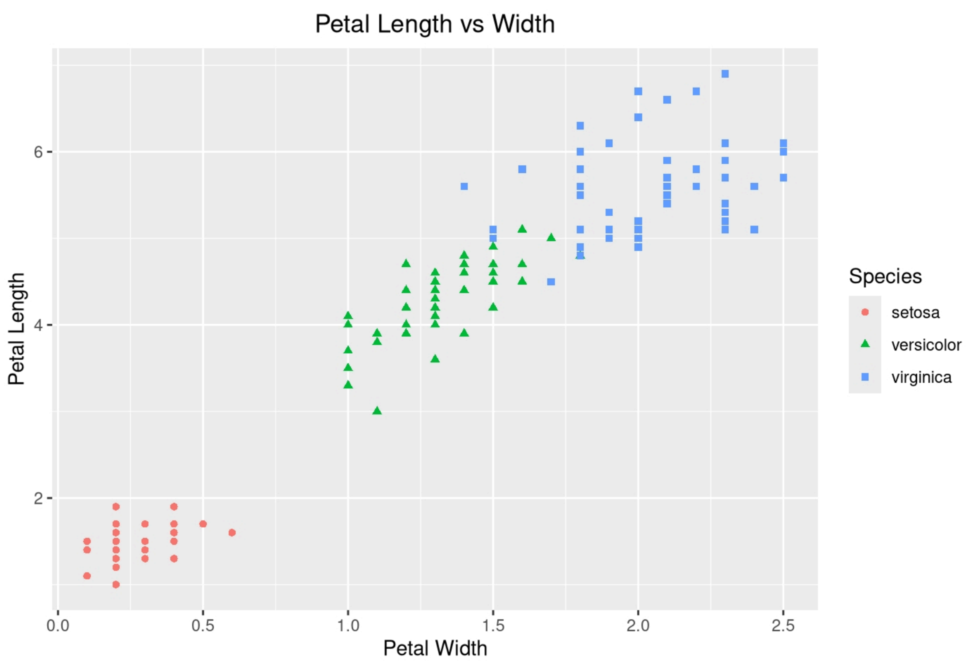
En el lenguaje de programación r existe un conjunto de datos llamados “iris” el cual tiene diversos datos de esta planta.

Pregunta a) Su objetivo es analizar dicho conjunto de datos y explicar la información contenida. Posteriormente, gráfica en R un gráfico que represente las características de los pétalos de la planta iris, en específico, su largo y el ancho del pétalo del conjunto “iris” el gráfico realizado debe ser como los vistos en clases.

Respuesta: El conjunto de datos “iris” contiene la información de la planta iris, en específico el ancho y largo de “sepal” (hoja de la flor) y “petal” (pétalos de la flor) y la especie de la planta. Estos datos se muestran en una tabla de 5 columnas y contiene 150 datos, dividido en 3 especies.

Se realiza un gráfico donde se relaciona el lago con el ancho del pétalo y se separa por forma las 3 especies, para ayudar en la visualización y comprensión de los datos.

Gráfica:



Código R:

```r
#================
# 3.- En el lenguaje de programación r existe un conjunto de datos llamados “iris”
#     el cual tiene diversos datos de esta planta.
#================

head(iris, 10)

#================
# a)  Su objetivo es analizar dicho conjunto de datos y explicar la información contenida. Posteriormente,
#     gráfica en R un gráfico que represente las características de los pétalos
#     de la planta iris, en específico, su largo y el ancho del pétalo del conjunto
#     “iris” el gráfico realizado debe ser como los vistos en clases.
#================

grafico = ggplot(iris, aes(x = Petal.Width, y = Petal.Length, colour = Species)) +
  geom_point(aes(shape = Species)) +
  xlab("Petal Width") + ylab("Petal Length") + ggtitle("Petal Length vs Width") +
  theme(plot.title = element_text(hjust = 0.5))
plot(grafico)
```

Consola R:

```
R 4.3.3 · ~/
> #================
> # 3.- En el lenguaje de programación r existe un conjunto de datos llamados “iris”
> #     el cual tiene diversos datos de esta planta.
> #================
>
> head(iris, 10)
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5          1.4         0.2  setosa
2           4.9         3.0          1.4         0.2  setosa
3           4.7         3.2          1.3         0.2  setosa
4           4.6         3.1          1.5         0.2  setosa
5           5.0         3.6          1.4         0.2  setosa
6           5.4         3.9          1.7         0.4  setosa
7           4.6         3.4          1.4         0.3  setosa
8           5.0         3.4          1.5         0.2  setosa
9           4.4         2.9          1.4         0.2  setosa
10          4.9         3.1          1.5         0.1  setosa
>
> #================
> # a)  Su objetivo es analizar dicho conjunto de datos y explicar la información contenida. Posteriormente,
> #     gráfica en R un gráfico que represente las características de los pétalos
> #     de la planta iris, en específico, su largo y el ancho del pétalo del conjunto
> #     “iris” el gráfico realizado debe ser como los vistos en clases.
> #================
>
> grafico = ggplot(iris, aes(x = Petal.Width, y = Petal.Length, colour = Species)) +
+   geom_point(aes(shape = Species)) +
+   xlab("Petal Width") + ylab("Petal Length") + ggtitle("Petal Length vs Width") +
+   theme(plot.title = element_text(hjust = 0.5))
> plot(grafico)
>
```

Pregunta b) Luego, ajuste el modelo lineal según lo visto en clases, obtenga las características del mismo, tales como residuos, coeficientes, entre otros, y finalmente grafique la regresión lineal simple.

Respuesta: Se aplica un modelo de regresión lineal simple, ya que una sola variable independiente (Petal Length) explica el comportamiento de la variable dependiente (Petal Width). Aplicado este modelo se obtienen los resultados que indican lo siguiente:

- Intercepto (c) = -0.3631
- pendiente (m) = 0.4158
- residuos “Residual standard error” (ϵ_i) = 0.2065

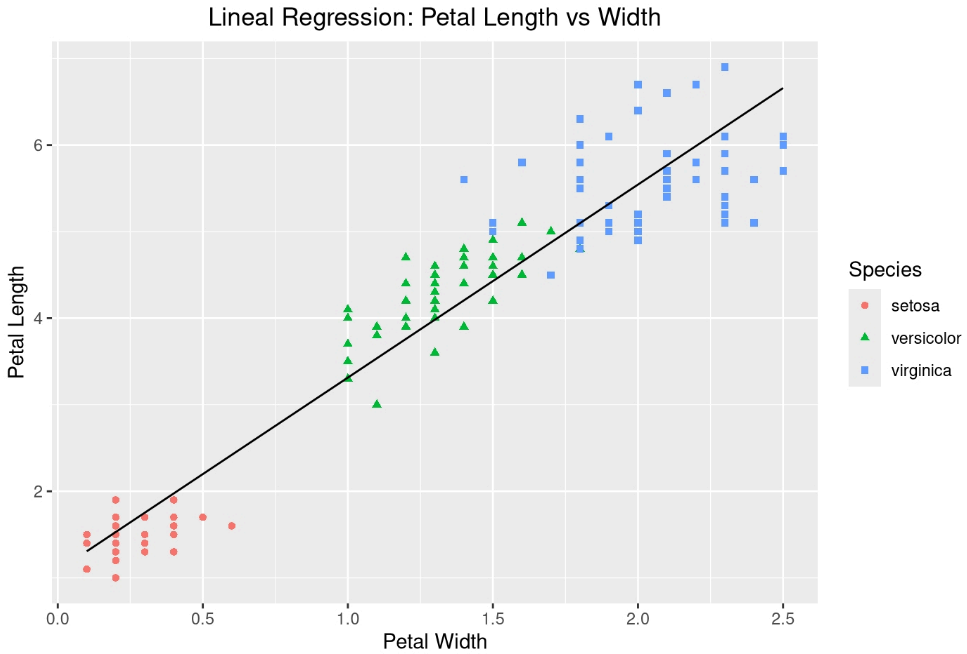
Por lo tanto, el modelo es el siguiente:

- $y_i = (mx_i * c) + \epsilon_i$ ⇒ (abajo)
- Petal Width = 0.4158 * Petal Length - 0.3631

De este modelo podemos decir que por cada unidad adicional en la longitud del pétalo, el ancho del pétalo aumenta en aproximadamente

0.4158 unidades, partiendo de un ancho del pétalo base de -0.3631 unidades cuando la longitud del pétalo es cero.

Gráfico:



Código R:

```r
#================
# b)  Luego, ajuste el modelo lineal según lo visto en clases, obtenga las características
#     del mismo, tales como residuos, coeficientes, entre otros, y finalmente grafique la
#     regresión lineal simple.
#================

# Calculando la regresion (Dependiente ~ Independiente)
regresion = lm(Petal.Width ~ Petal.Length, iris)

# Imprime datos de regresion
print(regresion)

# Imprime un resumen de la regresion
summary(regresion)

# Imprime el valor del r^2 del modelo
print(summary(regresion)$r.squared)

grafico = ggplot(iris, aes(x = Petal.Width, y = Petal.Length, colour = Species)) +
  geom_point(aes(shape = Species)) +
  xlab("Petal Width") + ylab("Petal Length") + ggtitle("Lineal Regresion: Petal Length vs Width") +
  geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="black", size=0.5, se = FALSE) +
  theme(plot.title = element_text(hjust = 0.5))
plot(grafico)
```

Consola R:

```
Console  Terminal  Background Jobs
R 4.3.3 · ~/
> #================
> # b)  Luego, ajuste el modelo lineal según lo visto en clases, obtenga las características
> #     del mismo, tales como residuos, coeficientes, entre otros, y finalmente grafique la
> #     regresión lineal simple.
> #================
>
> # Calculando la regresion (Dependiente ~ Independiente)
> regresion = lm(Petal.Width ~ Petal.Length, iris)
>
> # Imprime datos de regresion
> print(regresion)

Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)

Coefficients:
 (Intercept)  Petal.Length
     -0.3631        0.4158


>
> # Imprime un resumen de la regresion
> summary(regresion)

Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56515 -0.12358 -0.01898  0.13288  0.64272

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16

>
> # Imprime el valor del r^2 del modelo
> print(summary(regresion)$r.squared)
[1] 0.9271098
>
> grafico = ggplot(iris, aes(x = Petal.Width, y = Petal.Length, colour = Species)) +
+   geom_point(aes(shape = Species)) +
+   xlab("Petal Width") + ylab("Petal Length") + ggtitle("Lineal Regresion: Petal Length vs Width") +
+   geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="black", size=0.5, se = FALSE) +
+   theme(plot.title = element_text(hjust = 0.5))
> plot(grafico)
>
```

Pregunta c) Una vez realizado lo anterior, explique con sus palabras de forma clara lo que se observa en los gráficos, si los modelos son representativos o no, el significado de los valores obtenidos en R, etc.

Respuesta: La evaluación del modelo se hace con base en tres parámetros:

Parámetros	Explicación / Condición / Ho	Valor Obtenido
Coeficiente de determinación (r^2)	Este coeficiente indica que proporción de la variación de la variable independiente es explicada por las variables dependientes. En este caso se busca	0.9271098

	que la condición sea mayor a 0.95.	
p-value de la pendiente	Ho: La pendiente de la recta es cero	4.7e-16 ***
p-value del intercepto	Ho: Él interceptó de la recta es cero.	2e-16 ***

De los resultados obtenidos se determina lo siguiente:

1. Como el Coeficiente de determinación es un poco menos que el 0.95 esperado, se puede decir que el modelo responde regular al conjunto de datos evaluados, ya que no llega al nivel óptimo.
2. El p-value de la pendiente es menor que el nivel de significancia de 0,05, por lo tanto, se rechaza la Ho y podemos decir con un 95% de confianza que la pendiente de la recta no es cero.
3. El p-value del intercepto es menor que el nivel de significancia de 0,05, por lo tanto, se rechaza la Ho y podemos decir con un 95% de confianza que él interceptó de la recta no es cero.

Finalmente, se concluye que si bien el modelo no alcanza el óptimo (r^2 = 0.95) con el conjunto de datos analizados, pero su valor de 0.92 es muy cercano, por lo que el modelo responde de forma regular respecto al conjunto de datos. Él interceptó y la pendiente están bien calculadas y son representativas. En consecuencia, el modelo es correcto y está bien calculado y responde regular a los datos analizados.

Lo anterior se puede visualizar en el gráfico donde se ve que una parte de los datos está dentro del intervalo de confianza del modelo, pero hay muchos otros puntos que siguen alejados.

Gráfica:

Código R:

```
pep2_pregunta2.R   guia04_pregunta1.R   guia04_pregunta2.R   guia04_pregunta3.R
Source on Save                                    Run   Source
#===============
# c)  Una vez realizado lo anterior, explique con sus palabras de forma clara lo que se observa
#     en los gráficos, si los modelos son representativos o no, el significado de los valores
#     obtenidos en R, etc.
#===============

# Intervalo de confianza
confianza = confint(regresion, level=0.95)
print(confianza)

grafico = ggplot(iris, aes(x = Petal.Width, y = Petal.Length, colour = Species)) +
  geom_point(aes(shape = Species)) +
  xlab("Petal Width") + ylab("Petal Length") + ggtitle("Confidence Interval: Petal Length vs Width") +
  geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="black", size=0.5) +
  theme(plot.title = element_text(hjust = 0.5))
plot(grafico)

```

Consola R:

```
Console   Terminal   Background Jobs
R 4.3.3 · ~/
> #===============
> # c)  Una vez realizado lo anterior, explique con sus palabras de forma clara lo que se observa
> #     en los gráficos, si los modelos son representativos o no, el significado de los valores
> #     obtenidos en R, etc.
> #===============
>
> # Intervalo de confianza
> confianza = confint(regresion, level=0.95)
> print(confianza)
                  2.5 %     97.5 %
(Intercept)  -0.4416501 -0.2845010
Petal.Length  0.3968193  0.4346915
>
> grafico = ggplot(iris, aes(x = Petal.Width, y = Petal.Length, colour = Species)) +
+   geom_point(aes(shape = Species)) +
+   xlab("Petal Width") + ylab("Petal Length") + ggtitle("Confidence Interval: Petal Length vs Width") +
+   geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="black", size=0.5) +
+   theme(plot.title = element_text(hjust = 0.5))
> plot(grafico)
>
```