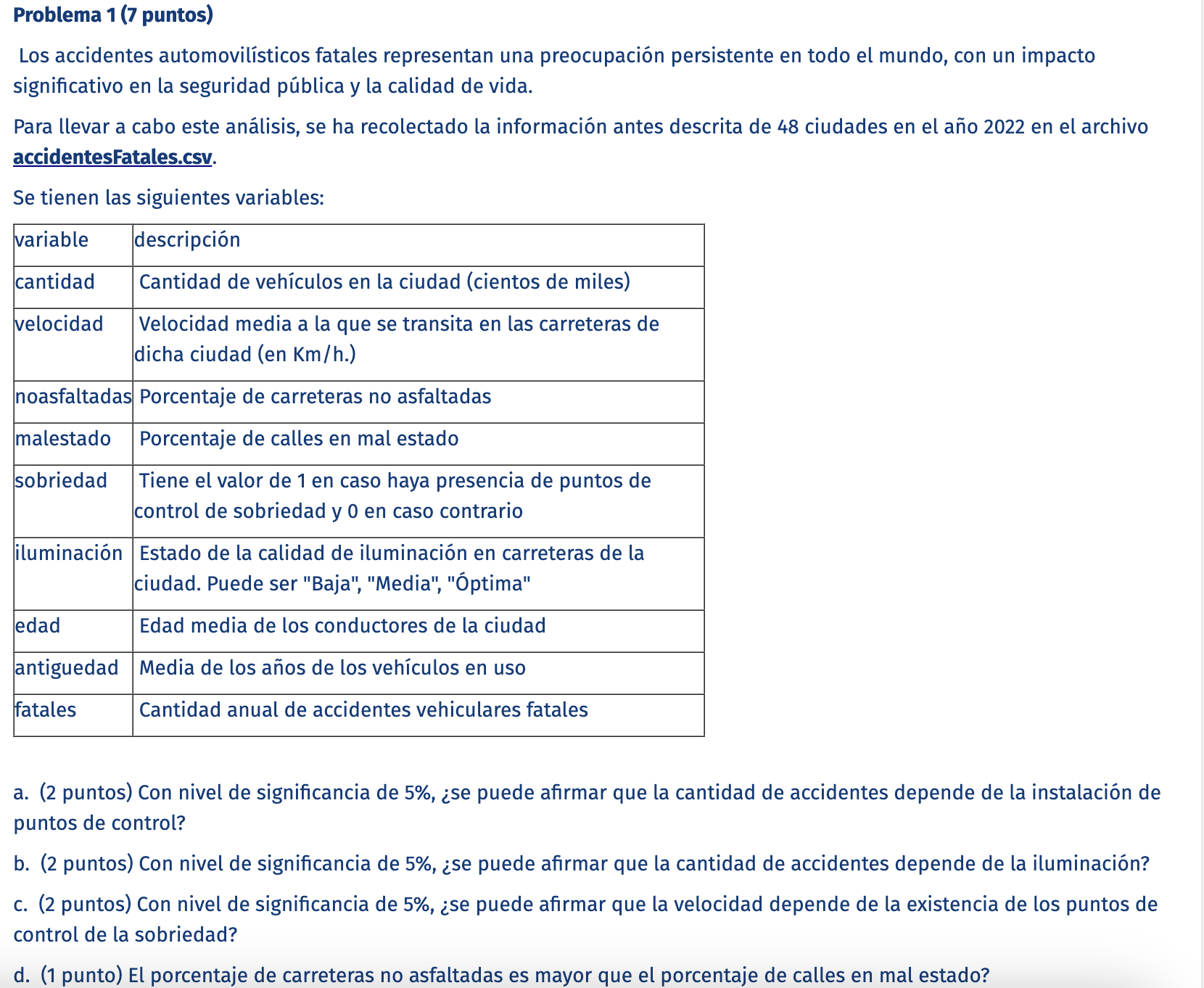
Examen 2023-2 Solución



1. a) **Revisar la normalidad**: Para cada grupo (ciudades con puntos de control y sin ellos), verifica si la variable "fatales" se distribuye normalmente. Esto puede hacerse con la prueba de Shapiro-Wilk.
2. **Comparación de medias**:
   * Si ambos grupos son normales, usa la prueba t de Student para muestras independientes.
   * Si los datos no son normales, utiliza la prueba de Mann-Whitney.
3. **Decidir basado en el p-valor**:
   * Si el p-valor es menor que 0.05, rechaza la hipótesis nula y concluye que hay una diferencia significativa en la cantidad de accidentes fatales entre ciudades con y sin puntos de control.
   * Si el p-valor es mayor o igual a 0.05, no rechazas la hipótesis nula, lo que significa que no hay evidencia suficiente para afirmar que los puntos de control influyen en la cantidad de accidentes fatales.
4. # Cargar los datos
5. datos <- read.csv("accidentesFatales.csv")
6. # Verificar normalidad para cada grupo
7. shapiro\_con <- shapiro.test(datos$fatales[datos$sobriedad == 1])
8. shapiro\_sin <- shapiro.test(datos$fatales[datos$sobriedad == 0])
9. # Decidir qué test usar basado en la normalidad
10. if (shapiro\_con$p.value > 0.05 && shapiro\_sin$p.value > 0.05) {
11. # Usar prueba t si ambos grupos son normales
12. t\_test\_result <- t.test(fatales ~ sobriedad, data = datos, var.equal = TRUE)
13. print(t\_test\_result)
14. } else {
15. # Usar Mann-Whitney si alguno de los grupos no es normal
16. mw\_test\_result <- wilcox.test(fatales ~ sobriedad, data = datos, exact = FALSE)
17. print(mw\_test\_result)
18. }

b) Para determinar si la cantidad de accidentes fatales depende del nivel de iluminación en las carreteras de las ciudades, podemos realizar un análisis de varianza (ANOVA), ya que estamos comparando más de dos grupos categorizados por la variable "iluminación" que tiene tres niveles (Baja, Media, Óptima).

El ANOVA es la técnica adecuada para comparar las medias de una variable cuantitativa (en este caso, la cantidad de accidentes fatales) entre varios grupos definidos por una variable categórica (el estado de la iluminación).

**Pasos para Realizar un ANOVA en R**

1. **Revisar la normalidad**: Antes de proceder con el ANOVA, es importante verificar que la distribución de la cantidad de accidentes en cada nivel de iluminación se aproxime a una distribución normal. Esto puede hacerse con la prueba de Shapiro-Wilk para cada grupo.
2. **Homogeneidad de varianzas**: También debes verificar que las varianzas entre los grupos sean homogéneas, lo cual es una suposición clave para el ANOVA. Esto puede evaluarse con la prueba de Levene.
3. **Realizar el ANOVA**: Si las suposiciones se cumplen, puedes proceder a realizar el ANOVA.
4. **Pruebas post hoc (si es necesario)**: Si el ANOVA muestra diferencias significativas, las pruebas post hoc como Tukey HSD ayudarán a determinar entre qué grupos específicos existen estas diferencias.
5. # Cargar los datos
6. datos <- read.csv("accidentesFatales.csv")
7. # Convertir la columna 'iluminacion' a factor
8. datos$iluminacion <- factor(datos$iluminacion, levels = c("Baja", "Media", "Optima"))
9. # Cargar dplyr y car si no están ya cargados
10. if (!require(dplyr)) {
11. install.packages("dplyr")
12. library(dplyr)
13. }
14. if (!require(car)) {
15. install.packages("car")
16. library(car)
17. }
18. # Verificar la normalidad de la distribución de 'fatales' para cada nivel de 'iluminacion'
19. normalidad <- datos %>%
20. group\_by(iluminacion) %>%
21. summarise(p\_value = shapiro.test(fatales)$p.value)
22. print(normalidad)
23. # Verificar homogeneidad de varianzas
24. levene\_result <- leveneTest(fatales ~ iluminacion, data = datos)
25. print(levene\_result)
26. # Realizar ANOVA
27. anova\_result <- aov(fatales ~ iluminacion, data = datos)
28. summary\_anova <- summary(anova\_result)
29. print(summary\_anova)
30. # Pruebas post hoc si el ANOVA es significativo
31. if (summary\_anova[[1]]$'Pr(>F)'[1] < 0.05) {
32. post\_hoc\_results <- TukeyHSD(anova\_result)
33. print(post\_hoc\_results)
34. }

Los resultados de tu análisis ANOVA y las pruebas post hoc proporcionan información valiosa sobre cómo el nivel de iluminación afecta la cantidad de accidentes fatales en las ciudades. Aquí te ofrezco una interpretación detallada de estos resultados:

**Resultados de la Prueba de Shapiro-Wilk**

* **Valores p de Shapiro-Wilk** para cada nivel de iluminación:
  + **Baja**: 0.183
  + **Media**: 0.140
  + **Óptima**: 0.961

Todos los valores p son mayores que 0.05, lo que indica que no hay evidencia suficiente para rechazar la hipótesis nula de normalidad. Esto sugiere que los datos de accidentes fatales para cada nivel de iluminación están distribuidos normalmente.

**Resultados de la Prueba de Levene**

* **Valor p de Levene**: 0.7269

Este valor p es también mayor que 0.05, indicando homogeneidad de las varianzas entre los grupos. Esta es una suposición clave para la validez de un ANOVA.

**Resultados del ANOVA**

* **Valor F**: 6.638
* **Valor p**: 0.00298

El valor p del ANOVA es menor que 0.05, indicando que hay diferencias estadísticamente significativas en las medias de accidentes fatales entre al menos dos niveles de iluminación.

**Resultados de las Pruebas Post Hoc (Tukey HSD)**

* **Diferencias entre los grupos**:
  + **Media-Baja**: La diferencia de medias es -450.36, con un intervalo de confianza que incluye cero (p-adj = 0.1685), indicando que no hay una diferencia estadísticamente significativa entre estos dos niveles.
  + **Óptima-Baja**: La diferencia de medias es -801.43, con un intervalo de confianza que no incluye cero (p-adj = 0.0038). Esto sugiere una diferencia estadísticamente significativa, con menos accidentes en el nivel Óptima en comparación con Baja.
  + **Óptima-Media**: La diferencia de medias es -351.07, con un intervalo de confianza que casi incluye cero (p-adj = 0.0882). No hay una diferencia estadísticamente significativa, pero el valor p ajustado es bastante bajo, indicando una tendencia que podría ser explorada con una muestra más grande.

**Conclusión**

La iluminación tiene un impacto significativo en la cantidad de accidentes fatales, particularmente entre los niveles de iluminación Baja y Óptima, donde la iluminación Óptima está asociada con significativamente menos accidentes. Esto sugiere que mejorar la iluminación en las carreteras podría ser una medida efectiva para reducir los accidentes fatales.

Estos resultados pueden informar a las autoridades y planificadores urbanos sobre la importancia de invertir en mejor iluminación en las carreteras como una estrategia para mejorar la seguridad vial y reducir los accidentes fatales. Si necesitas más detalles o tienes otras preguntas, estoy aquí para ayudarte.

1. c) **Verificar la normalidad de la distribución de la variable "velocidad"** en ambos grupos (con y sin puntos de control de sobriedad). Esto puede realizarse con la prueba de Shapiro-Wilk.
2. **Verificar la homogeneidad de varianzas** entre los dos grupos si los datos son normales, usando la prueba de Levene.
3. **Seleccionar y realizar la prueba estadística adecuada** basada en los resultados de las pruebas de normalidad y homogeneidad de varianzas.
4. **Interpretar el valor p** de la prueba seleccionada para tomar decisiones estadísticas.
5. # Cargar y verificar dplyr y car si no están ya cargados
6. if (!require(dplyr)) {
7. install.packages("dplyr")
8. library(dplyr)
9. }
10. if (!require(car)) {
11. install.packages("car")
12. library(car)
13. }
14. # Asegurarte de que 'sobriedad' es un factor
15. datos$sobriedad <- factor(datos$sobriedad, levels = c(0, 1))
16. # Verificar la normalidad para cada grupo de sobriedad
17. normalidad <- datos %>%
18. group\_by(sobriedad) %>%
19. summarise(p\_value = shapiro.test(velocidad)$p.value)
20. # Imprimir los resultados de normalidad nuevamente para claridad
21. print(normalidad)
22. # Verificar homogeneidad de varianzas con Levene's Test
23. levene\_result <- leveneTest(velocidad ~ sobriedad, data = datos)
24. print(levene\_result)
25. # Realizar la prueba apropiada basada en la normalidad de los datos
26. if (all(normalidad$p\_value > 0.05)) {
27. # Si todas las distribuciones son normales y las varianzas son homogéneas
28. if (levene\_result$p.value > 0.05) {
29. test\_result <- t.test(velocidad ~ sobriedad, data = datos, var.equal = TRUE)
30. } else {
31. test\_result <- t.test(velocidad ~ sobriedad, data = datos, var.equal = FALSE)
32. }
33. } else {
34. # Si alguna distribución no es normal, usa Mann-Whitney
35. test\_result <- wilcox.test(velocidad ~ sobriedad, data = datos, exact = FALSE)
36. }
37. # Imprimir los resultados de la prueba estadística elegida
38. print(test\_result)

**Interpretación de los Resultados de la Prueba de Mann-Whitney**

* **Estadístico W = 250.5**
* **Valor p = 0.6012**

Este valor p es bastante alto y excede el umbral típico de significancia de 0.05. Por lo tanto, no hay evidencia estadística suficiente para rechazar la hipótesis nula de que no hay diferencia en las medianas de velocidad entre las ciudades con puntos de control de sobriedad y las ciudades sin ellos.

**Conclusión**

Basado en este análisis, **no se puede afirmar con un nivel de significancia del 5% que la velocidad promedio dependa de la existencia de puntos de control de sobriedad**. Esto sugiere que la presencia de puntos de control de sobriedad no tiene un impacto estadísticamente significativo en la velocidad media de conducción en las ciudades.

Este resultado puede ser de interés para los planificadores urbanos y autoridades locales al considerar medidas para influir en los comportamientos de conducción. Aunque los puntos de control de sobriedad pueden tener otros beneficios, como reducir la incidencia de conducción bajo la influencia del alcohol, este análisis específico no muestra una relación directa con cambios en la velocidad media de conducción.

Si necesitas explorar otros aspectos de los datos, realizar análisis adicionales o tienes más preguntas sobre este tema, estoy aquí para ayudarte.

diferencia <- datos$noasfaltadas - datos$malestado

shapiro\_test <- shapiro.test(diferencia)

print(shapiro\_test)

# Realizar la prueba t de muestras pareadas o Wilcoxon según la normalidad

if (shapiro\_test$p.value > 0.05) {

# Los datos de la diferencia son normales, usar prueba t

t\_test\_result <- t.test(diferencia, mu = 0, alternative = "greater")

print(t\_test\_result)

} else {

# Los datos de la diferencia no son normales, usar Wilcoxon

wilcoxon\_test <- wilcox.test(diferencia, mu = 0, alternative = "greater", paired = TRUE)

print(wilcoxon\_test)

}

Gracias por proporcionar los resultados de tu análisis. Aquí está la interpretación de lo que encontraste con la prueba t de muestras pareadas para la diferencia entre el porcentaje de carreteras no asfaltadas y el porcentaje de calles en mal estado:

**Resultados del Test de Shapiro-Wilk**

* **W = 0.97001**
* **p-value = 0.2535**

El valor p del test de Shapiro-Wilk es mayor que 0.05, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula de que los datos de la diferencia son normalmente distribuidos. Esto valida el uso de la prueba t de muestras pareadas para tu análisis.

**Resultados de la Prueba t de Muestras Pareadas**

* **Estadístico t = 0.14179**
* **Grados de libertad (df) = 47**
* **p-value = 0.4439**
* **Intervalo de confianza del 95%**: [-6.093923, Infinito]
* **Estimación de la media de la diferencia = 0.5625**

**Interpretación**

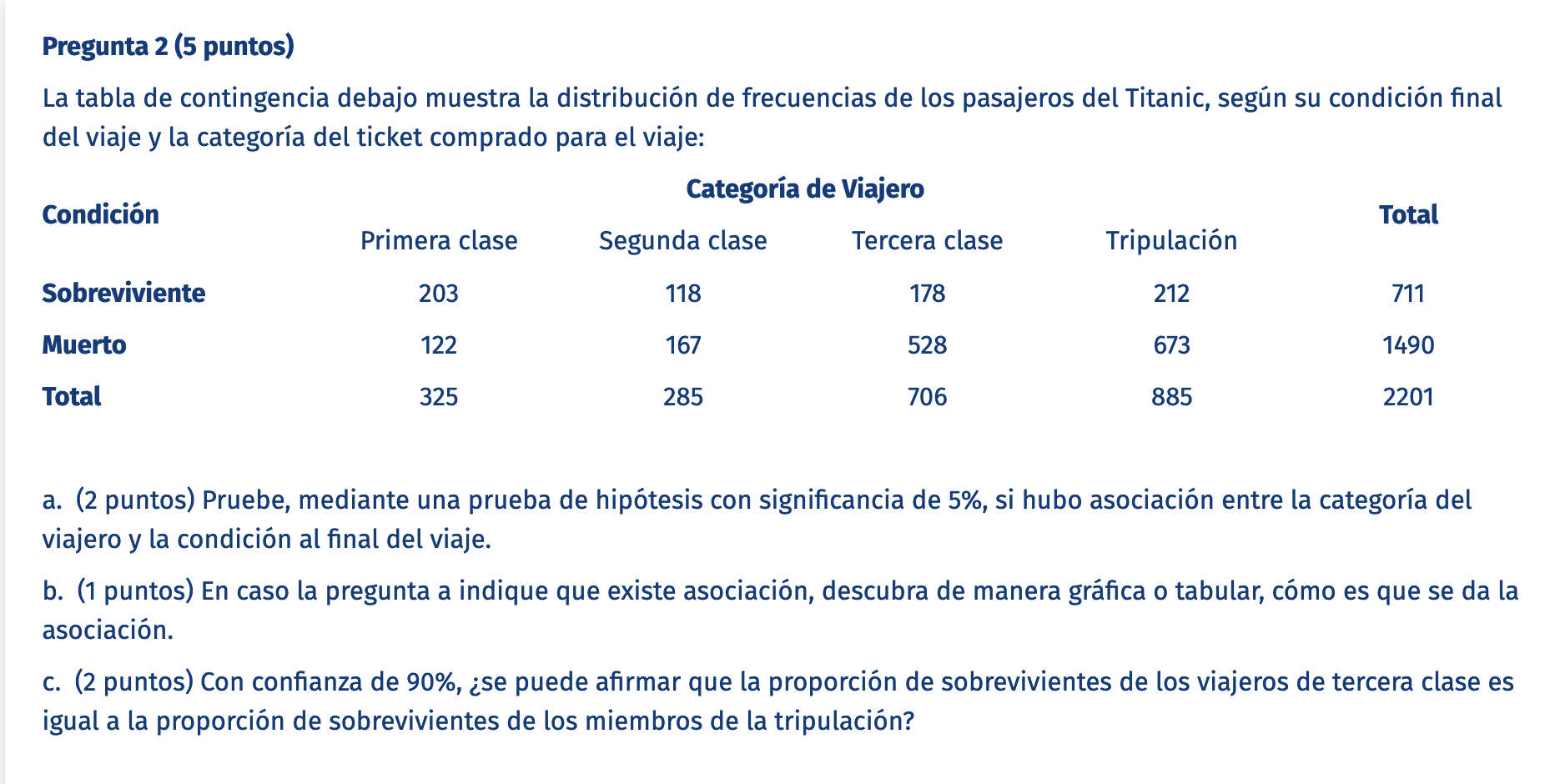
El valor p de la prueba t es 0.4439, lo cual es bastante superior al nivel de significancia estándar de 0.05. Esto significa que no hay evidencia estadística suficiente para afirmar que el porcentaje de carreteras no asfaltadas es mayor que el porcentaje de calles en mal estado, al menos no de manera que sea estadísticamente significativa.

El intervalo de confianza para la diferencia media incluye cero y se extiende hacia el infinito positivo, lo que indica que, aunque la media de la diferencia es positiva (0.5625), no es estadísticamente significativa desde el punto de vista de este test.

**Conclusión**

Con base en los resultados de la prueba, no se puede afirmar, con un nivel de confianza del 95%, que el porcentaje de carreteras no asfaltadas es mayor que el porcentaje de calles en mal estado. Esto sugiere que no hay suficiente diferencia entre estas dos variables como para ser consideradas estadísticamente diferentes en esta muestra de datos.

Si necesitas explorar otros aspectos de los datos o realizar análisis adicionales, o si tienes alguna pregunta más, estaré aquí para ayudarte.



a)

* **H0 (Hipótesis nula):** No hay asociación entre la categoría del viajero y la condición al final del viaje.
* **H1 (Hipótesis alternativa):** Hay una asociación entre la categoría del viajero y la condición al final del viaje.

# Crea la tabla de contingencia basada en los datos proporcionados

observados <- matrix(c(203, 122, 118, 167, 178, 528, 212, 673), nrow = 2,

byrow = TRUE, dimnames = list(Condicion = c("Sobreviviente", "Muerto"),

Categoria = c("Primera clase", "Segunda clase", "Tercera clase", "Tripulación")))

# Realizar la prueba chi-cuadrado

chi2\_test <- chisq.test(observados)

# Imprimir los resultados de la prueba chi-cuadrado

print(chi2\_test)

El valor-p extremadamente bajo (menor que 2.2e-16) indica que puedes rechazar la hipótesis nula con mucha confianza. Esto significa que **existe una asociación estadísticamente significativa entre la categoría del viajero y su condición final (sobreviviente o muerto) al final del viaje**. Los datos sugieren que la probabilidad de sobrevivir o morir en el desastre no era la misma para todas las categorías de pasajeros y tripulación.

b) Dado que se encontró una asociación significativa, procedemos con la visualización. Crearemos un gráfico de barras para mostrar visualmente las diferencias en las proporciones de sobrevivientes y muertos entre las diferentes categorías de viajeros. Este tipo de visualización ayuda a entender de manera más clara cómo las diferencias entre los grupos contribuyen a la asociación detectada.

# Reestructurar la matriz para asegurarse de que los nombres se asignen correctamente

observados <- matrix(c(203, 118, 178, 212, 122, 167, 528, 673), ncol = 4, byrow = TRUE,

dimnames = list(c("Sobreviviente", "Muerto"),

c("Primera clase", "Segunda clase", "Tercera clase", "Tripulación")))

# Graficar los resultados utilizando un gráfico de barras

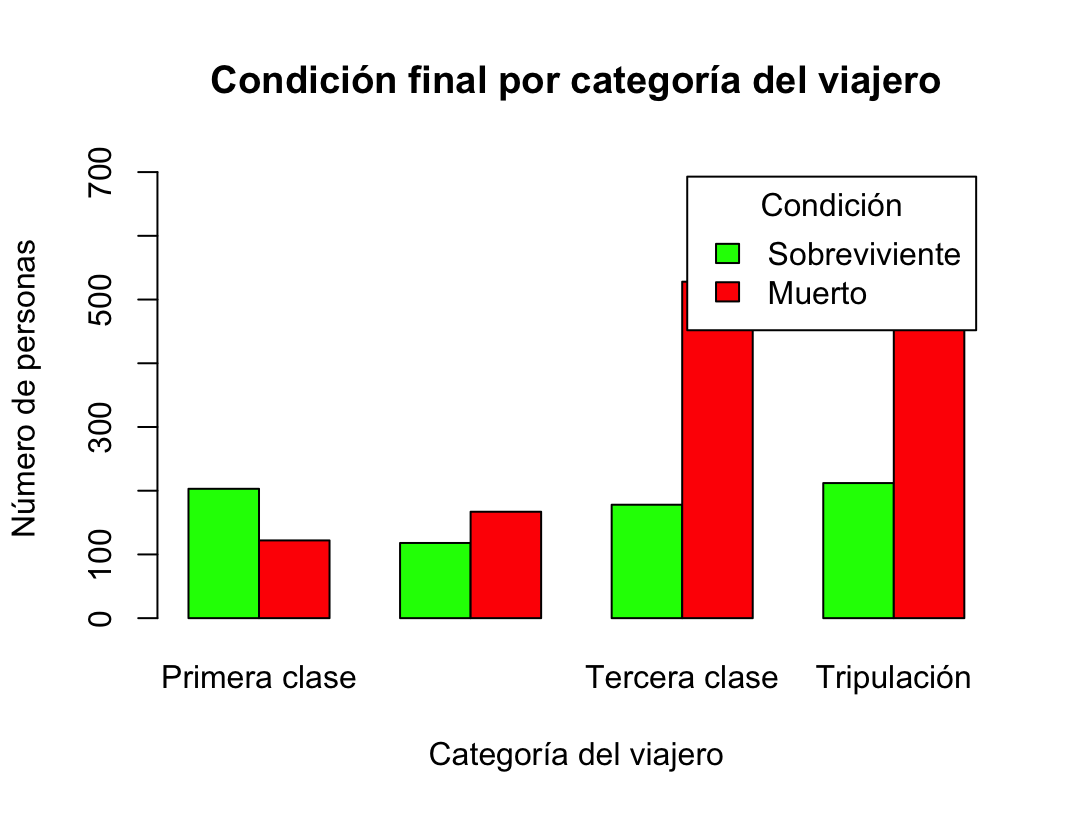
barplot(observados, beside = TRUE, col = c("green", "red"),

legend.text = TRUE, args.legend = list(title = "Condición", fill = c("green", "red")),

ylim = c(0, max(observados) + 50),

main = "Condición final por categoría del viajero",

ylab = "Número de personas", xlab = "Categoría del viajero")



* **Primera Clase:** La diferencia entre sobrevivientes y muertos es menos marcada, con una mayor proporción de sobrevivientes.
* **Segunda Clase:** También presenta un número considerable de sobrevivientes, aunque los muertos superan ligeramente a los sobrevivientes.
* **Tercera Clase:** Muestra una gran diferencia, con muchos más muertos que sobrevivientes, lo que indica una baja tasa de supervivencia.
* **Tripulación:** Al igual que en la tercera clase, la cantidad de muertos es significativamente mayor que la de sobrevivientes.

c) necesitamos comparar la proporción de sobrevivientes entre los viajeros de tercera clase y los miembros de la tripulación para determinar si son estadísticamente iguales, con un nivel de confianza del 90%.

Usaremos el intervalo de confianza para la diferencia entre dos proporciones:

# Asumiendo que los datos de sobrevivientes y total para tercera clase y tripulación están correctos

n\_sobrevivientes\_tercera <- 178

n\_total\_tercera <- 706

n\_sobrevivientes\_tripulacion <- 212

n\_total\_tripulacion <- 885

# Usar prop.test para comparar proporciones con un intervalo de confianza del 90%

resultado\_prop <- prop.test(x = c(n\_sobrevivientes\_tercera, n\_sobrevivientes\_tripulacion),

n = c(n\_total\_tercera, n\_total\_tripulacion),

conf.level = 0.90, correct = FALSE)

# Imprimir los resultados de la prueba

print(resultado\_prop)

2-sample test for equality of proportions without continuity correction

data: c(n\_sobrevivientes\_tercera, n\_sobrevivientes\_tripulacion) out of c(n\_total\_tercera, n\_total\_tripulacion)

X-squared = 0.33569, df = 1, p-value = 0.5623

alternative hypothesis: two.sided

90 percent confidence interval:

-0.02319336 0.04834661

sample estimates:

prop 1 prop 2

0.2521246 0.2395480

1. **Valor p: 0.5623**
   * Este valor p indica que no hay una diferencia estadísticamente significativa entre las proporciones de sobrevivientes de la tercera clase y de la tripulación. Con un valor p mayor que el nivel de significancia típico (en este caso, 0.10 para un intervalo de confianza del 90%), no rechazamos la hipótesis nula de que las proporciones son iguales.
2. **Intervalo de Confianza del 90%: [-0.02319336, 0.04834661]**
   * Este intervalo de confianza incluye el valor cero, lo que significa que la diferencia entre las proporciones de sobrevivientes de la tercera clase y de la tripulación no es estadísticamente significativa al nivel de confianza del 90%. En otras palabras, con un 90% de confianza, no podemos afirmar que las proporciones sean diferentes.
3. **Estimaciones de Proporción:**
   * **Proporción de la Tercera Clase (prop 1): 0.2521**
   * **Proporción de la Tripulación (prop 2): 0.2395**
   * Las proporciones estimadas muestran que la tercera clase tiene una ligera ventaja en términos de proporción de sobrevivientes en comparación con la tripulación, pero esta diferencia no es significativa.

### Conclusión

Con un nivel de confianza del 90%, no hay suficiente evidencia para afirmar que la proporción de sobrevivientes de los viajeros de tercera clase es diferente de la proporción de sobrevivientes de los miembros de la tripulación. Esto podría indicar que ambos grupos tuvieron oportunidades similares de sobrevivir, o que otros factores no examinados en este análisis podrían influir en las tasas de supervivencia de ambos grupos.

Este resultado es importante para entender la dinámica de supervivencia en situaciones de desastre y podría ser relevante para estudios históricos, investigaciones sobre la seguridad de los viajes, o políticas relacionadas con la gestión de emergencias.

