www.nataliejulian.com
Natalie Julian

Anova

Beer goggles effect

El conjunto de datos goggles (de la librería *WRS2*) trata sobre los efectos del alcohol en la selección de pareja en clubes nocturnos. Hay 48 participantes: 24 hombres y 24 mujeres. El investigador llevó a 3 grupos de 8 participantes a un club nocturno. Un grupo recibió cerveza sin alcohol (placebo), un grupo recibió 2 pintas (concentración media de alcohol en 500 ml de cerveza) y un grupo recibió 4 pintas (concentración alta de alcohol en 500 ml de cerveza). Al final de la velada, el investigador tomó una fotografía de la persona con la que el participante estaba charlando. El atractivo de la persona en la foto fue luego evaluado por jueces de manera independiente, formando un indicador en una escala del 0 al 100 del atractivo del acompañante.

La hipótesis es que después de consumir alcohol, los umbrales de percepción del atractivo disminuyen, por lo que se quiere evaluar dependiendo del consumo de alcohol, el cómo se distribuye el indicador del atractivo del acompañante.

Análisis previo

a) Obtenga tablas y gráficos de manera de generar intuiciones para responder la pregunta de interés. Comente.

Respuesta

Nuestro factor de interés estudiar es el alcohol, el cual toma los valores *None*, para representar al placebo, *2 Pints*, que consumieron con concentración media de alcohol y *4 Pints*, que consumieron con concentración alta de alcohol. Es decir, la variable factor alcohol posee r $= 3$ niveles. Considere i el subíndice para indicar los niveles del factor.

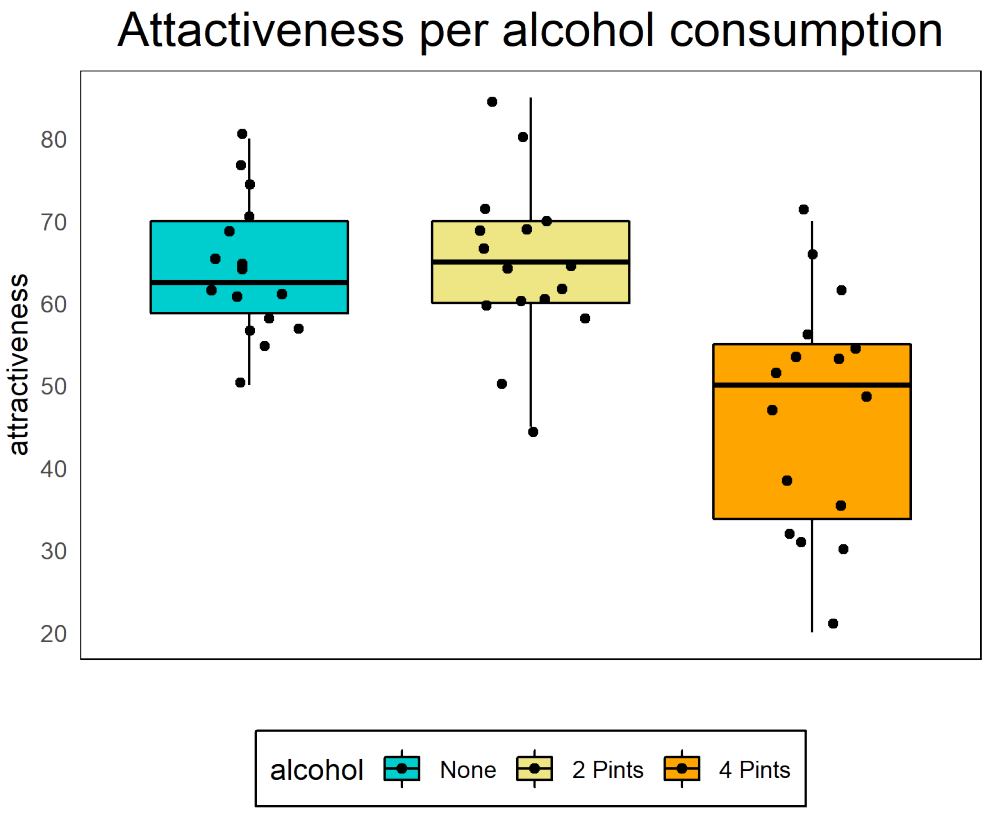
```
dim(goggles)
[1] 48  3

table(goggles$alcohol)

None 2 Pints 4 Pints
 16      16      16
```

Podemos observar que estamos en un caso balanceado, es decir, hay la misma cantidad de individuos en cada categoría del factor. También podemos usar el paquete dplyr para confirmar lo anterior.

```
#Definamos la información de los niveles del factor
r <- length(unique(alcohol)) #son 3 niveles del factor

#En este caso, podemos ver que el estudio es balanceado
library(dplyr)
Caso_Balanceado <- goggles %>%
  group_by(alcohol) %>%
  summarize(n_i = length(attractiveness)) #son iguales
ni <- Caso_Balanceado$n_i[1]

#Además, n_T = \sum_{i = 1}^r n_i
nT <- sum(Caso_Balanceado$n_i) #48
```

Además, obtenemos las medias por nivel del factor:

```
#Los valores \mu_i.gorrito = \Bar{Y}_{i.}
Medias_alcohol <- goggles %>%
  group_by(alcohol) %>%
  summarize(mu_i = mean(attractiveness)) %>%
  mutate(dif = mu_i - mean(attractiveness))

#Gráfico Boxplot

df<-data.frame(goggles)

library(ggplot2)
library(ggpubr)

(p<-ggboxplot(df, y="attractiveness", x="alcohol", fill="alcohol", add = "jitter")+
    xlab("")+
    ylab("attractiveness")+
    scale_y_continuous(breaks = round(seq(min(df$attractiveness), max(df$attractiveness), by = 10),0))+
    ggtitle("Attactiveness per alcohol consumption")+
    scale_fill_manual(values=c("cyan3", "khaki2", "orange"))+
    theme_minimal()+
    theme(legend.position="bottom", axis.text.x = element_blank(), plot.title=element_text(hjust=0.5,
          panel.background = element_rect(fill = "transparent"),
          plot.background = element_rect(fill = "transparent", color = NA),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.background = element_rect(fill = "transparent"),
          legend.box.background = element_rect(fill = "transparent")))

#ggsave(p, filename = "myplot.png",  bg = "transparent")
```




```
#Se puede observar que las personas que las personas
#que no consumen alcohol o consumen con bajas concentraciones
#de este, poseen parejas en el club mejor evaluadas en terminos
#de attractiveness, en comparación a las personas que tienen una
```

```
#alta concentración de alcohol.
```

Modelamiento

b) Plantee el modelo de medias de celda, ajuste el modelo y realice el test correspondiente. Concluya.

Respuesta
El modelo de medias de celda es:

$$\text{Modelo 1} \quad Y_{ij} = \mu_i + \epsilon_{ij}$$
$$\epsilon_{ij} \text{ iid } N(0, \sigma^2)$$

donde la variable Y_ij es el atractivo y mu_i son las medias por nivel del factor, es decir, las medias de los grados de alcohol.

Recordemos el modelo $Y_{ij} = \mu_i + \epsilon_{ij}$, con $\epsilon_{ij} \sim N(0, \sigma^2)$, con $i = 1, ..., r; j = 1, ..., n_i$. Para ajustar el modelo de anova, utilizamos el modelo aov con el factor alcohol que explique la variable atractivo.

```
modelo <- aov(attractiveness ~ alcohol)
```

Test de igualdad de medias:

$$H_0 : \text{Las medias por grupo son todas iguales} \iff \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \text{Las medias por grupo no son todas iguales} \iff \exists(i, i') \quad \mu_i \neq \mu_{i'}$$

Tenemos dos formas de rechazar:

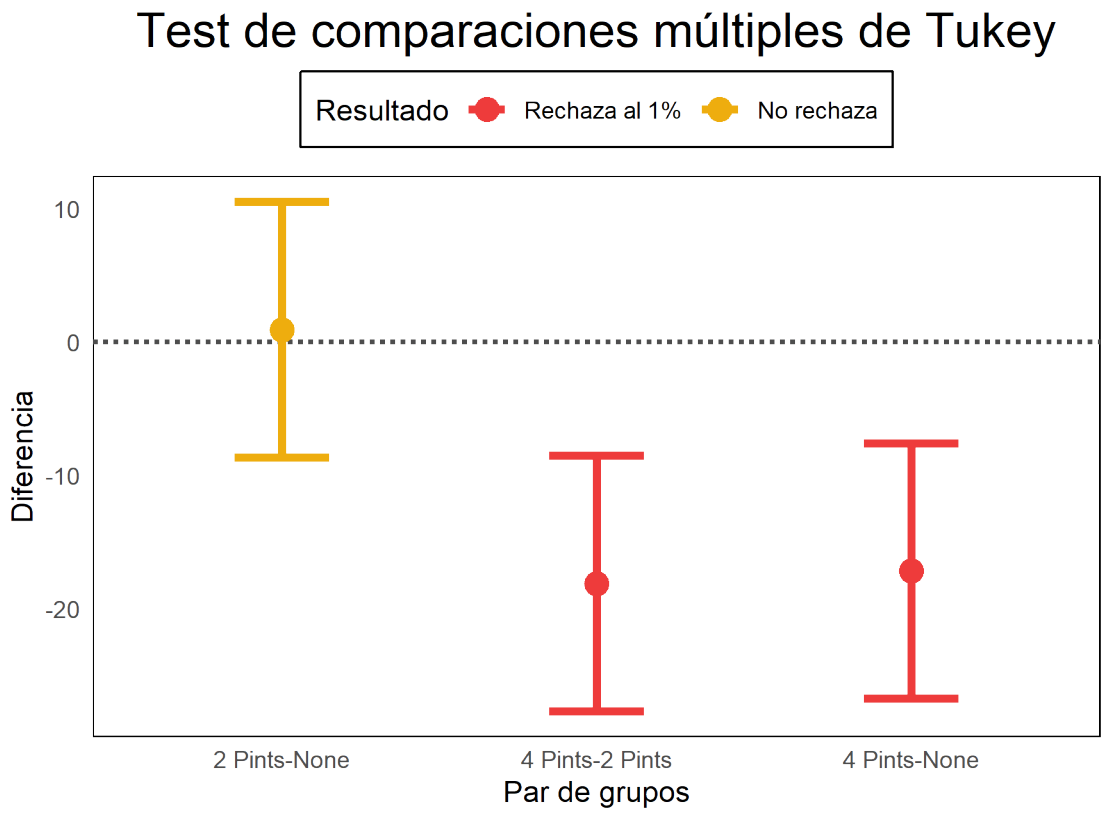
```
#######Método 1 de rechazo#######
valor_p <- anova(modelo)[1,5] #2.882911e-05
#Rechazo H_0 si valor-p < \alpha = 0.05
valor_p < 0.05 #TRUE
#Vemos que el valor-p es 2.882911e-05, por lo que
#rechazamos H_0 con un nivel de significancia del 5%,
#por lo tanto hay alguna media, por
#lo menos que difiere de las demás.

#######Método 2 de rechazo#######
#Usando un nivel de significancia del 5%
qf(0.95, r - 1, nT - r) #Cuantil con el que se compara en el
#summary o anova del modelo:
F_0 <- anova(modelo)[1,4] #Estadístico = 13.30699
#Rechazo H_0 si
F_0 > qf(0.95, r - 1, nT - r) #TRUE
```

Dado el test anterior, rechazamos la igualdad de media con un nivel de significancia del 5%, es decir, rechazamos que los atractivos medios de las parejas elegidas en el club nocturno por las personas con diferente grado de alcohol en la sangre es igual.

c) Determine cuáles son los grupos que se diferencian entre sí (y cuáles no), utilice el test correspondiente y comente.

```
##########Comparaciones de a pares.
#Si hemos detectado diferencias significativas
#entre las medias de las poblaciones. ¿Sería
#posible saber cuáles son los grupos que generan
#estas diferencias?

####Tukey:
intervals <- TukeyHSD(modelo)
intervals

tky <-as.data.frame(TukeyHSD(modelo)$alcohol)
tky$pair <- rownames(tky)
```

```
(p2<-ggplot(tky, aes(colour=cut(`p adj`, c(0, 0.01, 0.05, 1),
                               label=c("Rechaza al 1%","Rechaza al 5%","No rechaza")))) +
    geom_hline(yintercept=0, lty="11", colour="grey30", lwd=1) +
    geom_errorbar(aes(pair, ymin=lwr, ymax=upr), width=0.3, lwd=1.5) +
    geom_point(aes(pair, diff), size=4) +
    labs(colour="Resultado")+
    xlab("Par de grupos")+
    ylab("Diferencia")+
    ggtitle("Test de comparaciones múltiples")+
    scale_colour_manual(values=c("brown2", "darkgoldenrod2", "chartreuse3"))+
    theme_minimal()+
    theme(legend.position="top", plot.title=element_text(hjust=0.5, size=18),
          panel.background = element_rect(fill = "transparent"),
          plot.background = element_rect(fill = "transparent", color = NA),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.background = element_rect(fill = "transparent"),
          legend.box.background = element_rect(fill = "transparent"))
)
```



Podemos ver que las personas sin alcohol o con concentración media de alcohol en la sangre encuentran parejas igualmente atractivas, ya que el IC de la diferencia de las medias de 2 pintas y sin alcohol (None), contiene el valor 0.

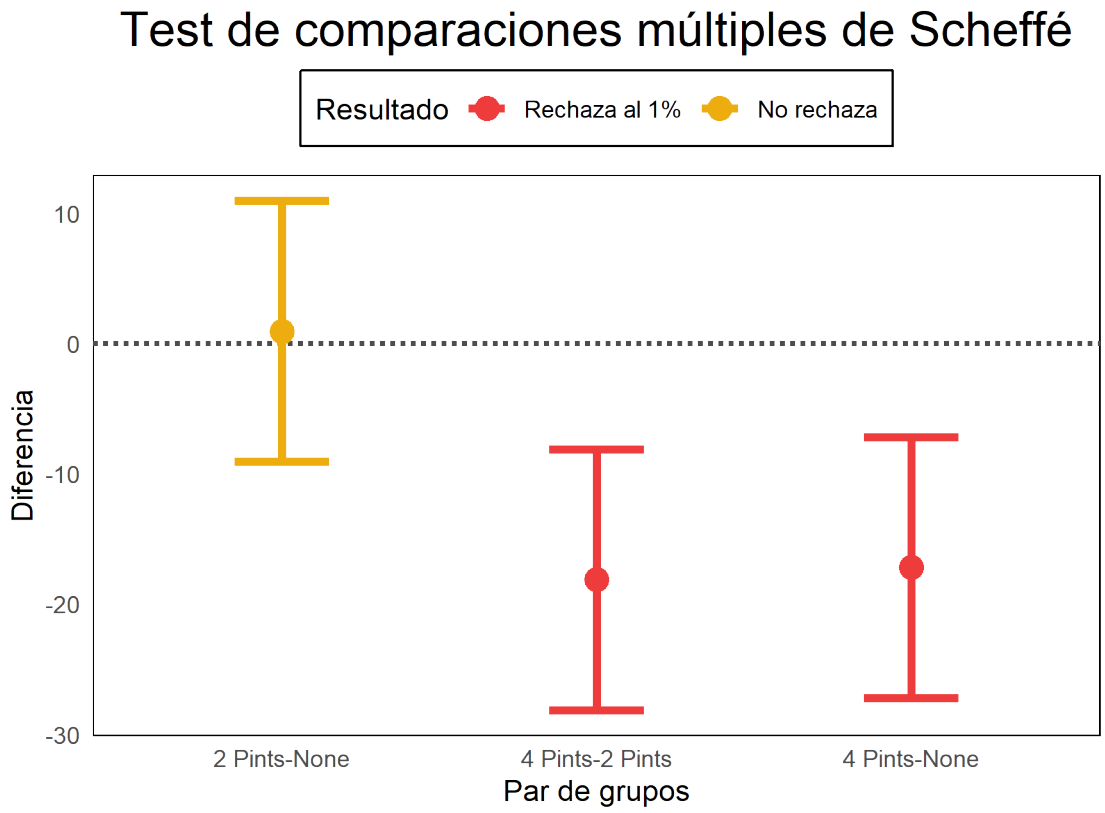
También podemos ver IC($\mu_3 - \mu_1$), en donde μ_3 es el atractivo medio de las parejas de personas con alto grado de alcohol en la sangre y μ_1 es el atractivo medio de las parejas de personas sin alcohol en la sangre, contiene solo valores menores que 0.

Por lo anterior, podemos decir que además de diferir los atractivos medios de las parejas que encuentran las personas que tienen alto nivel de alcohol en la sangre con respecto a las personas que no tienen alcohol en la sangre, vemos que el atractivo medio de las parejas que encuentran las personas con alto grado de alcohol es menor que el atractivo medio de las parejas que encuentran las personas sin alcohol.

Lo mismo se puede observar cuando comparamos a las personas que tienen alto nivel de alcohol en la sangre con aquellas que tienen concentración de alcohol medio.

También podemos realizar los test de Shceffé y Bonferroni para las comparaciones de a pares:

```
#Scheffé
library(DescTools)
ScheffeTest(modelo)

scheffe <-as.data.frame(ScheffeTest(modelo)$alcohol)
scheffe$pair <- rownames(scheffe)


(p3<-ggplot(scheffe, aes(colour=cut(‘pval‘, c(0, 0.01, 0.05, 1),
                                label=c("Rechaza al 1%","Rechaza al 5%","No rechaza")))) +
    geom_hline(yintercept=0, lty="11", colour="grey30", lwd=1) +
    geom_errorbar(aes(pair, ymin=lwr.ci, ymax=upr.ci), width=0.3, lwd=1.5) +
    geom_point(aes(pair, diff), size=4) +
    labs(colour="Resultado")+
    xlab("Par de grupos")+
    ylab("Diferencia")+
    ggtitle("Test de comparaciones múltiples de Scheffé")+
    scale_colour_manual(values=c("brown2", "darkgoldenrod2", "chartreuse3"))+
    theme_minimal()+
    theme(legend.position="top", plot.title=element_text(hjust=0.5, size=18),
          panel.background = element_rect(fill = "transparent"),
          plot.background = element_rect(fill = "transparent", color = NA),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.background = element_rect(fill = "transparent"),
          legend.box.background = element_rect(fill = "transparent"))
)

#ggsave(p3, filename = "myplot3.png",  bg = "transparent")
#Donde se obtienen conclusiones similares al anterior.
```




```
#Bonferroni
library(multcomp)

### Matriz de comparaciones:
```

```
K <- rbind(c(1, -1, 0),
           c(1, 0, -1),
           c(0, 1, -1))

modelo.gh <- glht(modelo, linfct = mcp(alcohol = K))

confint(modelo.gh)

# Valores p con método de bonferroni
summary(modelo.gh, test = adjusted("bonferroni"))

#Donde se obtienen conclusiones similares al anterior.
```

Donde se obtienen conclusiones similares al anterior.

d) Se desea comparar el atractivo medio de las parejas planteando las siguientes hipótesis conjuntas:

- ¿Serán el atractivo medio de las parejas de las personas sin alcohol en la sangre, igual al atractivo medio de las parejas de las personas que tienen alta concentración de alcohol en la sangre?
- ¿Será el promedio de los atractivos medios de las personas sin alcohol y con concentración media de alcohol en la sangre igual al atractivo medio de las parejas de las personas que tienen alta concentración de alcohol en la sangre?

Realice los test de manera conjunta.

Respuesta

El test de hipótesis de manera conjunta es:

$$H_{01} : \mu_1 = \mu_3$$

$$H_{02} : \frac{\mu_1 + \mu_2}{2} = \mu_3$$

$$H_1 : \text{al menos una de los dos combinaciones no es igual}$$

La hipótesis nula se puede traducir como:

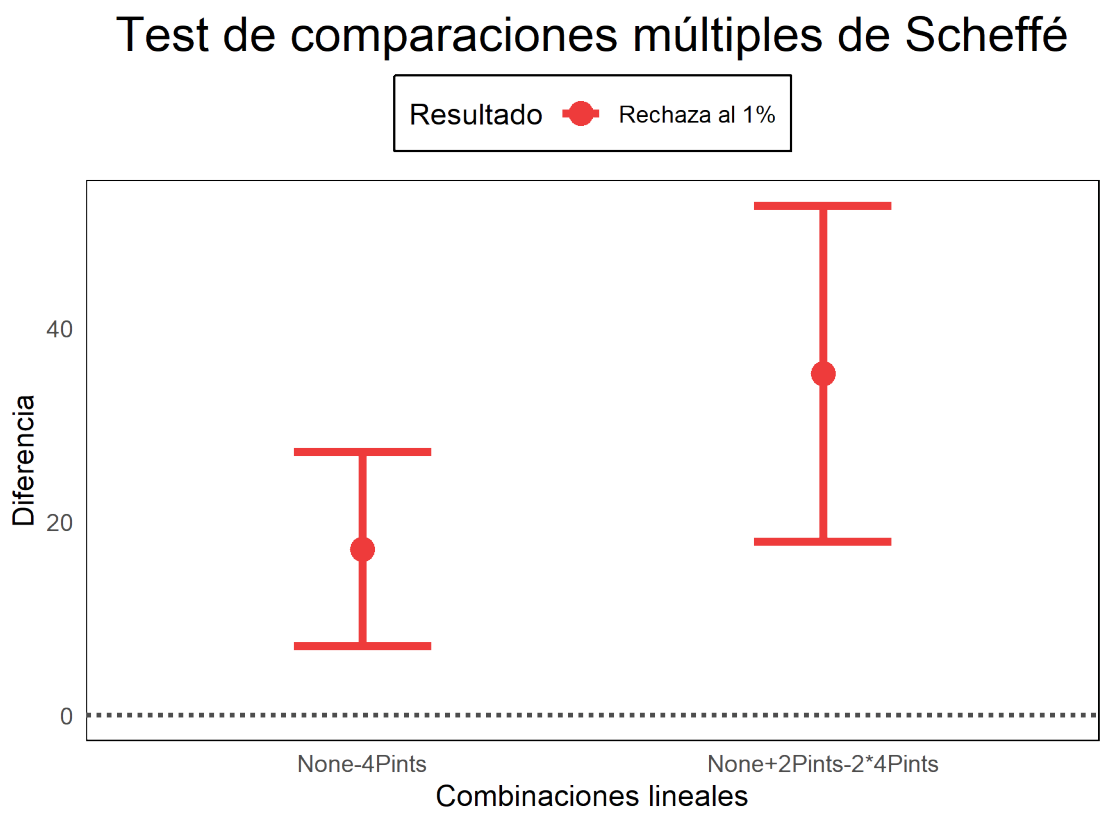
$$H_0 : \mu_1 - \mu_3 = 0 \quad \text{y} \quad \mu_1 + \mu_2 - 2\mu_3 = 0$$

Usando la matriz:

$$H_0 : \begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & -2 \end{pmatrix} \times \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

```
#Scheffé
scheffe2<-as.data.frame(ScheffeTest(modelo, contrasts = matrix(c(1,0,-1,1,1,-2),ncol=2,nrow=3))$al

scheffe2$pair <- rownames(scheffe2)

(p4<-ggplot(scheffe2, aes(colour=cut(‘pval‘, c(0, 0.01, 0.05, 1),
                                     label=c("Rechaza al 1%","Rechaza al 5%","No rechaza")))) +
    geom_hline(yintercept=0, lty="11", colour="grey30", lwd=1) +
    geom_errorbar(aes(pair, ymin=lwr.ci, ymax=upr.ci), width=0.3, lwd=1.5) +
    geom_point(aes(pair, diff), size=4) +
    scale_x_discrete(labels=c("None-4Pints", "None+2Pints-2*4Pints"))+
    labs(colour="Resultado")+
    xlab("Combinaciones lineales")+
    ylab("Diferencia")+
    ggtitle("Test de comparaciones múltiples de Scheffé")+
```

```
    scale_colour_manual(values=c("brown2", "darkgoldenrod2", "chartreuse3"))+
    theme_minimal()+
    theme(legend.position="top", plot.title=element_text(hjust=0.5, size=18),
          panel.background = element_rect(fill = "transparent"),
          plot.background = element_rect(fill = "transparent", color = NA),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.background = element_rect(fill = "transparent"),
          legend.box.background = element_rect(fill = "transparent"))
)

#ggsave(p4, filename = "myplot4.png",  bg = "transparent")
```



Para escribir el contraste en R, debemos escribir la matriz traspuesta de

$$\begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & -2 \end{pmatrix}$$

Pues la matriz debe ser de $r \times c$, donde r es el número de niveles (3) y c, el número de contrastes (2), en donde la suma de los coeficientes de cada constraste debe sumar 0.

```
#Rechazamos cada combinación por separado con un nivel de significancia
#del 5%, pero el test de manera conjunta se rechaza, es decir, rechazamos
#H_0 con un nivel de significancia mayor al 5%.

#Bonferroni
###contrastes:
K2 <- rbind(c(1, 0, -1),
            c(1, 1, -2))
modelo.gh <- glht(modelo, linfct = mcp(alcohol = K2))
confint(modelo.gh)
# Bonferroni corrected p-values
summary(modelo.gh, test = adjusted("bonferroni"))
#Tenemos los valores-p de cada combinación lineal, en donde ambos
#valores-p son muy pequeños (menores que alpha), por lo que
```

```
#en ambas combinaciones rechazamos H_0 a un nivel de significancia del 5%
#para ambos test.
#Es decir, rechazamos que el atractivo medio de las parejas de las personas
#sin alcohol es igual al atractivo medio de las parejas de las
#personas con alto grado de alcohol en la sangre.
#Además, rechazamos que el promedio de los atractivos medios de las
#parejas que son elegidas por personas sin alcohol en la sangre y con
#una concentración media de alcohol en la sangre sea igual al atractivo
#medio de las parejas de las personas que tienen alta concentración
#de alcohol en la sangre.
```

Trampas atractivas por su color

Estamos interesados en conocer si hay colores más atractivos para los insectos. Para ello se diseñaron trampas con los siguientes colores: **amarillo, azul, blanco y verde** y se cuantificó el número de insectos que quedaban atrapados.

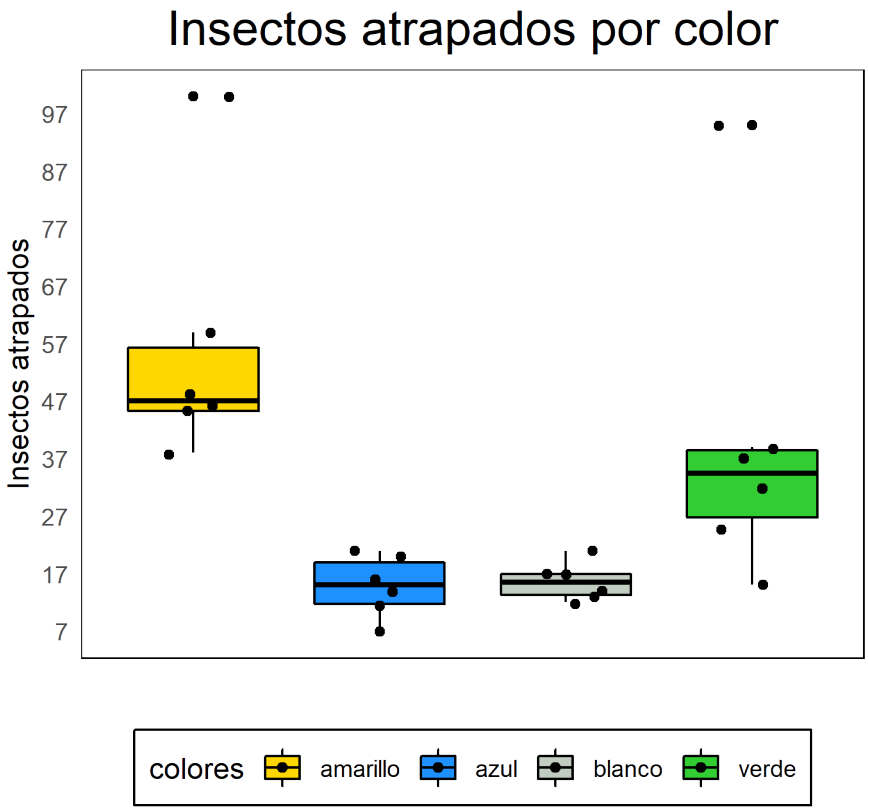
```
insectos <- c(16,11,20,21,14,7,37,32,15,25,39,95,21,12,14,17,13,17,45,59,48,46,38,100)
colores <- as.factor(c(rep(c("azul", "verde", "blanco", "amarillo"), each = 6)))
```

Análisis previo

a) Realice una comparación inicial del número medio de insectos que quedaron atrapados por color y un boxplot para comparar los números de insectos atrapados en las trampas de colores. Entregue un análisis exploratorio.

Respuesta

```
#Vamos a obtener las medias por tratamiento, es decir, vamos a obtener
#los valores \mu_i.gorrito = \Bar{Y}_{i.}
library(dplyr)
#Modelo medias de celdas
Medias_colores <- atractivo %>%
  group_by(colores) %>%
  summarize(mu_i = mean(insectos))

#Gráfico
df<-data.frame(atractivo)

(p5<-ggboxplot(df, y="insectos", x="colores", fill="colores", add = "jitter")+
    xlab("")+
    ylab("Insectos atrapados")+
    scale_y_continuous(breaks = round(seq(min(df$insectos), max(df$insectos), by = 10),0))+
    ggtitle("Insectos atrapados por color")+
    scale_fill_manual(values=c("gold1", "dodgerblue1", "honeydew3", "limegreen"))+
    theme_minimal()+
    theme(legend.position="bottom", axis.text.x = element_blank(), plot.title=element_text(hjust=0
          panel.background = element_rect(fill = "transparent"),
          plot.background = element_rect(fill = "transparent", color = NA),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.background = element_rect(fill = "transparent"),
          legend.box.background = element_rect(fill = "transparent")))

#ggsave(p5, filename = "myplot5.png",  bg = "transparent")
```




```
#Análisis exploratorio:
#Pareciera que los colores más atractivos para los insectos son el amarillo
#y el verde, en comparación a los colores azul y blanco.
```

Modelamiento

b) Plantee el modelo a utilizar. Realice el test F de significancia.

Respuesta

```
#Vamos a ajustar el modelo anova,
#con el factor colores que explique la variable
#número de insectos que quedan atrapados en la trampa
fm = aov(insectos ~ colores)

#Recordemos el modelo Y_ij = \mu_i + \epsilon_ij,
#con \epsilon_ij \sim N(0, \sigma^2),
#con i = 1,..., r; j = 1,..., n_i
valor_p <- anova(fm)[1,5] #0.001803896
#Rechazamos H_0 si valor-p < 0.05
valor_p < 0.05 #TRUE
#Rechazamos H_0, por lo tanto hay alguna media, por
#lo menos que difiere de las demás.
```

c) Realice la validación del modelo.

Respuesta

Para realizar la validación del modelo, debemos corroborrar que se cumpla la independencia, la normalidad y la homocedasticidad.

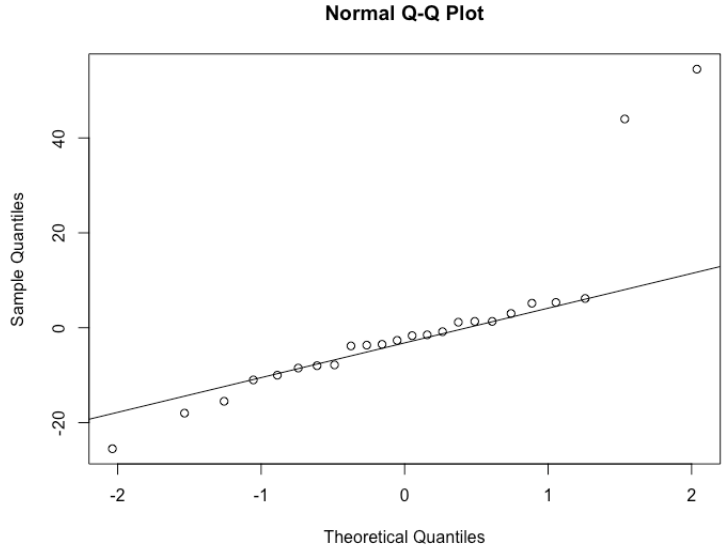
- Independencia: Test de Durbin Watson

```
#H_0: la correlación de los errores es 0
#H_1: la correlación de los errores es mayor que 0
```

```
library(lmtest)
dwtest(fm, alternative =  "two.sided")
#Valor-p = 0.1037
#Rechazamos H_0 si valor-p < \alpha
0.1037 < 0.05 #FALSE
#Por lo tanto no rechazamos el supuesto.
```

– Normalidad: Gráfico y test de Shapiro-Wilks

```
qqnorm(fm$residuals)
qqline(fm$residuals)
```




```
#El test de Shapiro-Wilk:
#H_0: normalidad versus H_1: no normalidad
shapiro.test(fm$residuals) #p-value = 7.207e-05
#El valor-p es muy pequeño, en particular, menor que 0,
#por lo que rechazamos H_0, por lo tanto rechazamos la
#normalidad.
```

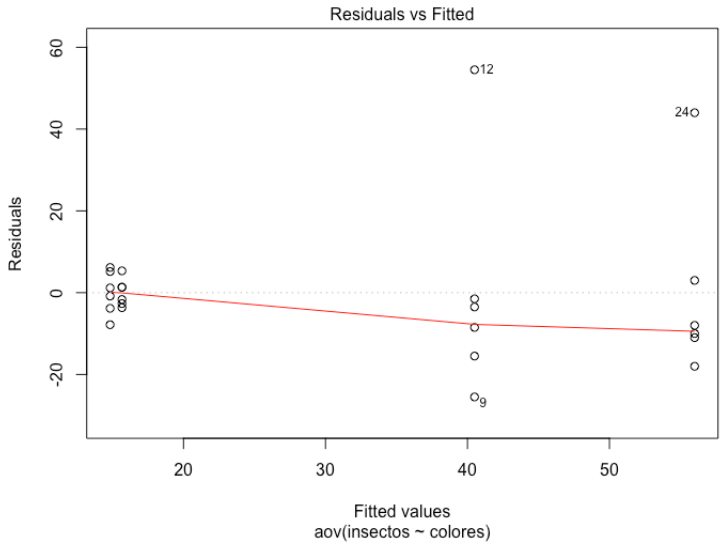
– Homocedasticidad: test de Hartley

```
#Test de Hartley:
#H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4
#H_1: Heterocedasticidad

#Comparando la desviación máxima con la mínima
#obtenemos una orientación sobre la falta de
#homocedasticidad
library(PMCMRplus)
hartleyTest(insectos ~ colores, data = atractivo) #p-value = 0.001039

#Rechazamos H_0 si valor-p < 0.05
0.001039 < 0.05 #TRUE, por lo rechazamos la homocedasticidad
```

d) ¿Por qué no se estarán cumpliendo todos los supuestos?

Respuesta

```
#Si volvemos a ver el gráfico del modelo:
plot(fm)
#Podemos identificar como observaciones más alejadas del resto,
#estas son la 12 y la 24.
```




```
#Si es que vemos la base de datos:
atractivo[c(12,24),] #vemos que son mayores a las demás observaciones, identificando
dos posibles outliers.
```