

Tarea diplomado: Analisis de caso: Mortalidad en MarineFjordland en 2021

Carolina San Martin

2022-06-30

Titulo: “Analisis de caso, mortalidad en MarineFjordland en 2021”

Descripción del problema a resolver

En 2021, en Marinefjordland, un zona de fiordos con 5 areas de manejo (simil a un ‘barrio’ en Chile) donde se cultiva salmón del atlantico, se vieron afectados 87 de 150 centros de cultivo. El problema fue una mortalidad significativa de inicio repentino, comenzando en un nivel bajo, pero aumentando rápidamente hasta el 5% de la población en riesgo por día en los sitios afectados. Los datos fueron recopilados en 2021, sin resultados concluyentes, no fue posible determinar el patogeno que lo causó, ni aislarlo. Por lo tanto la definición de caso fue: mortalidad significativa de inicio súbito que comienza en un nivel bajo, pero aumenta rápidamente al 5% o más de la población en riesgo por día en los sitios afectados. El centro de cultivo se considera caso (1) cuando las mortalidades alcanzan el 5% por día y no caso (0). Se probaron varios enfoques terapéuticos, pero ninguno ayudo significativamente a evitar la progresión de la enfermedad (aumento de mortalidad). Se cosecharon algunos de los sitios afectados, en la medida que la talla se los permitió, 5 sitios sacrificaron el stock, y en el resto, continuaron su etapa de engorda. En aquellos sitios que continuaron creciendo las pérdidas de peces fueron: Mín. 15 %, Máx. 65 %, y Media 45 %. Los datos fueron recopilados por la autoridad sanitaria nacional competente a principios de junio de 2021, de 150 granjas. 87 de estos fueron identificados como casos en ese momento. Los datos disponibles de la encuesta incluyen:

- SiteName (Nombre centro)
- ManagementArea (Area de Manejo)
- Case (Caso 1/0)
- MeanWaterTemperature (Temperatura promedio)
- Density (Densidad)
- Company (Compañía)
- Vaccine (Tipo de vacuna)
- SeaLice (Presencia de Sealice)
- GillDisease (Enfermedad branquial)
- FailedSmolt (Calidad smolt)

Descripción de las variables en estudio, factores a analizar y numero total de observaciones.

Se realizara un analisis exploratorio a la base de datos de 150 centros de cultivo, con su correspondientes informacion

```
D <- read_delim("MarineFjordland.txt")
```

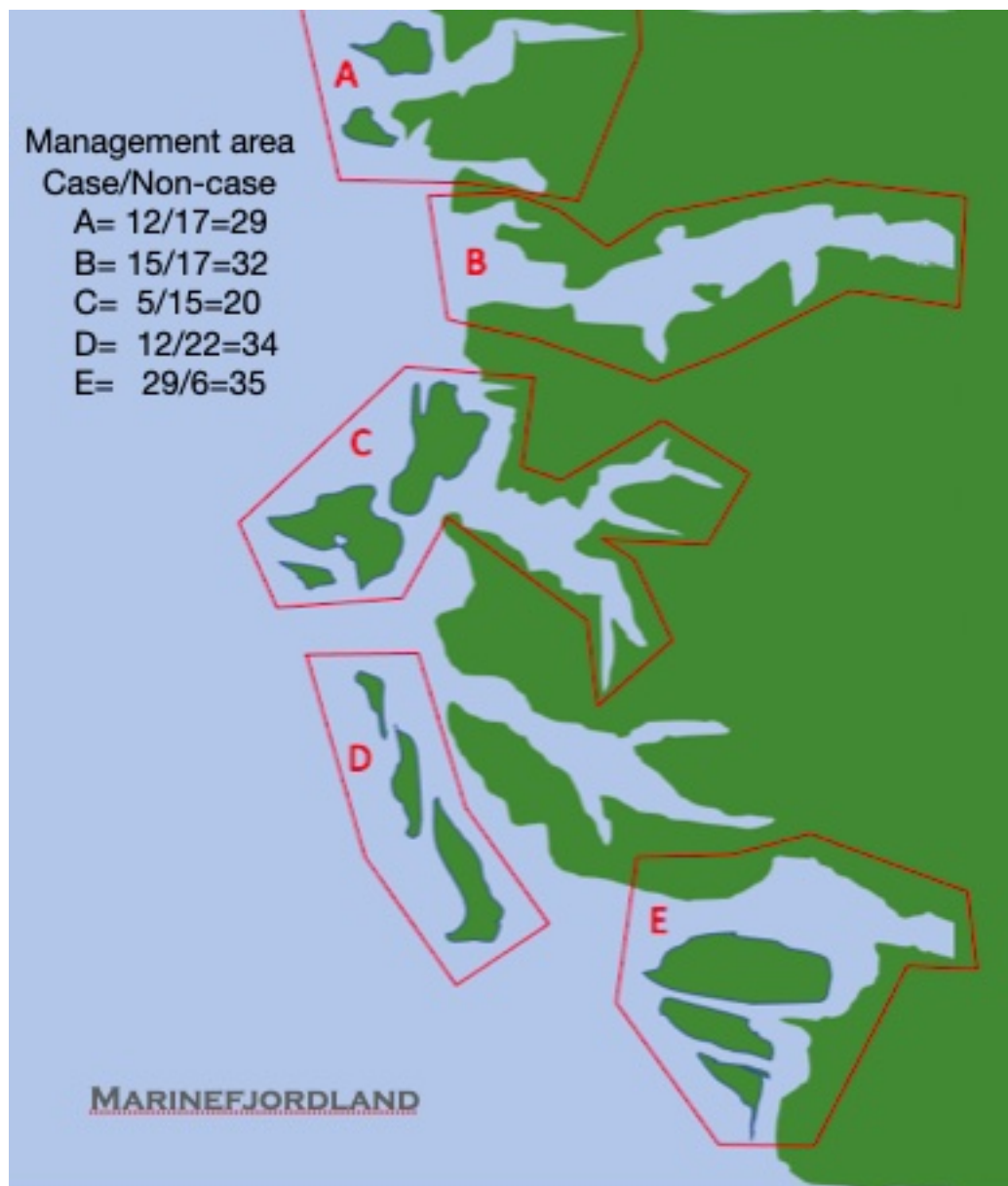


Figure 1: Figura 1: Areas de Manejo A, B, C, D y E de la zona MarineFjordland

```
## New names:
## Rows: 150 Columns: 11
## -- Column specification
## ----- Delimiter: "\t" chr
## (4): SiteName, ManagementArea, Company, Vaccine dbl (6): Case,
## MeanWaterTemperature, Density, SeaLice, GillDisease, FailedSmolt lgl (1): ...11
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...11`
```

Preguntas Iniciales

¿Cuántas y qué tipo de variables se dispone para el análisis? 8 variables: Area de manejo (cualitativa nominal), Compañía (cualitativa nominal), Tipo de vacuna (cualitativa nominal), Temperatura promedio (cuantitativo continuo), Densidad (cuantitativo continuo), Presencia de Sealice (variable aleatoria discreta con distribucion bernoulli), enfermedad branquial (variable aleatoria discreta con distribucion bernoulli), calidad de smolt (variable aleatoria discreta con distribucion bernoulli). ¿Cuales son los tratamientos? Caso 1 (positivo, es decir mortalidad mayor al 5% diario), Caso 0 (negativo, es decir no se mortalidad mayor al 5% diario), variable aleatoria discreta con distribucion bernoulli ¿La base de datos está completa?, ¿tiene errores? ¿Es posible responder las causas de mortalidad con los datos disponibles? ¿la cantidad de datos y variables permite hacer un análisis estadístico?

Resumen y Visualización

```
resumen<-D%>%group_by(Case)%>%summarise(N=n(), mean(Density), Variance= var(Density))
kable(resumen)
```

Case	N	mean(Density)	Variance
0	77	10.53896	21.04195
1	73	14.09205	38.33200

Centros de Cultivo por area de Manejo (Barrio)

```
table(D$ManagementArea)
```

```
##
##  A  B  C  D  E
## 29 32 20 34 35
```

Casos Totales 0=negativo 1= caso positivo

```
table(D$Case, D$ManagementArea)
```

```
##
##      A  B  C  D  E
## 0 17 17 15 22  6
## 1 12 15  5 12 29
```

Histograma Temperatura promedio

```
ggplot(D, aes(x=D$MeanWaterTemperature))+  
  geom_histogram(color="black", fill="grey", bins = 10)+theme_classic()+theme(text = element_text(size=
```

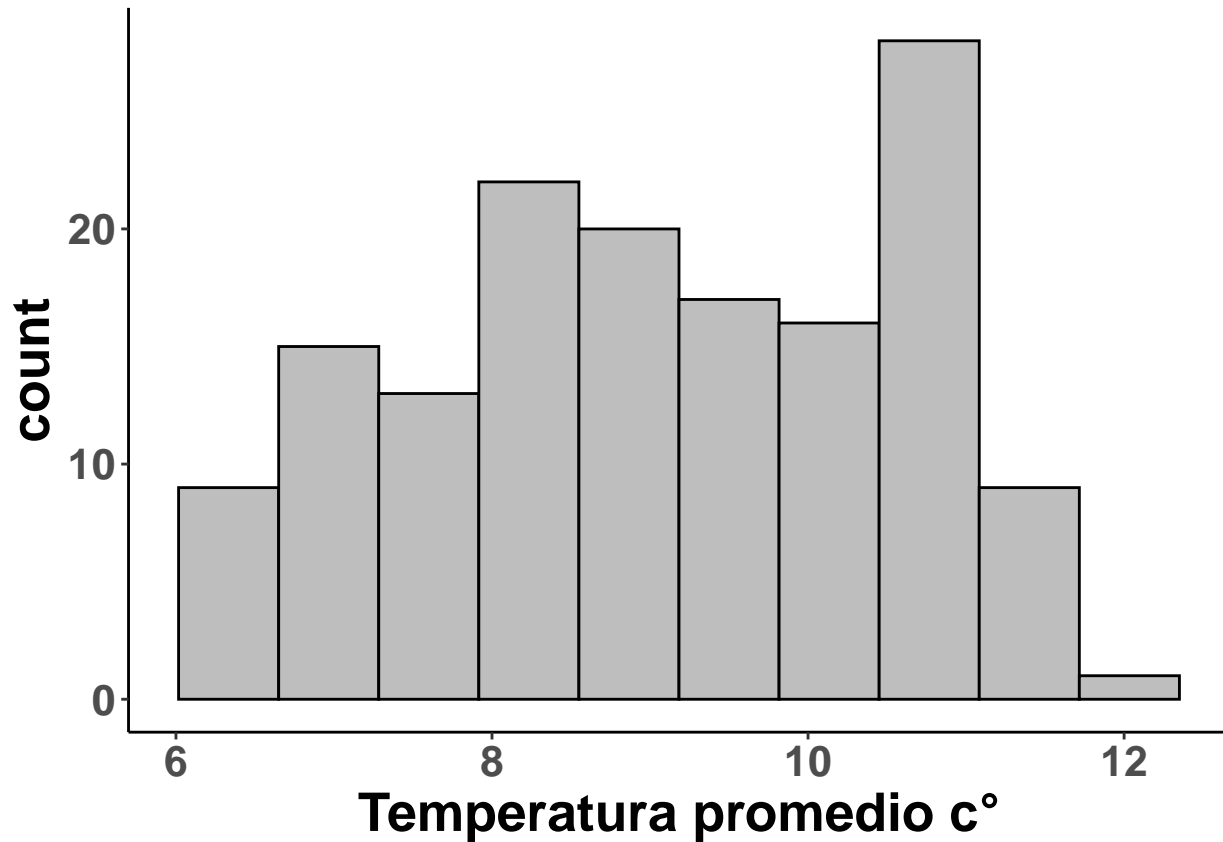
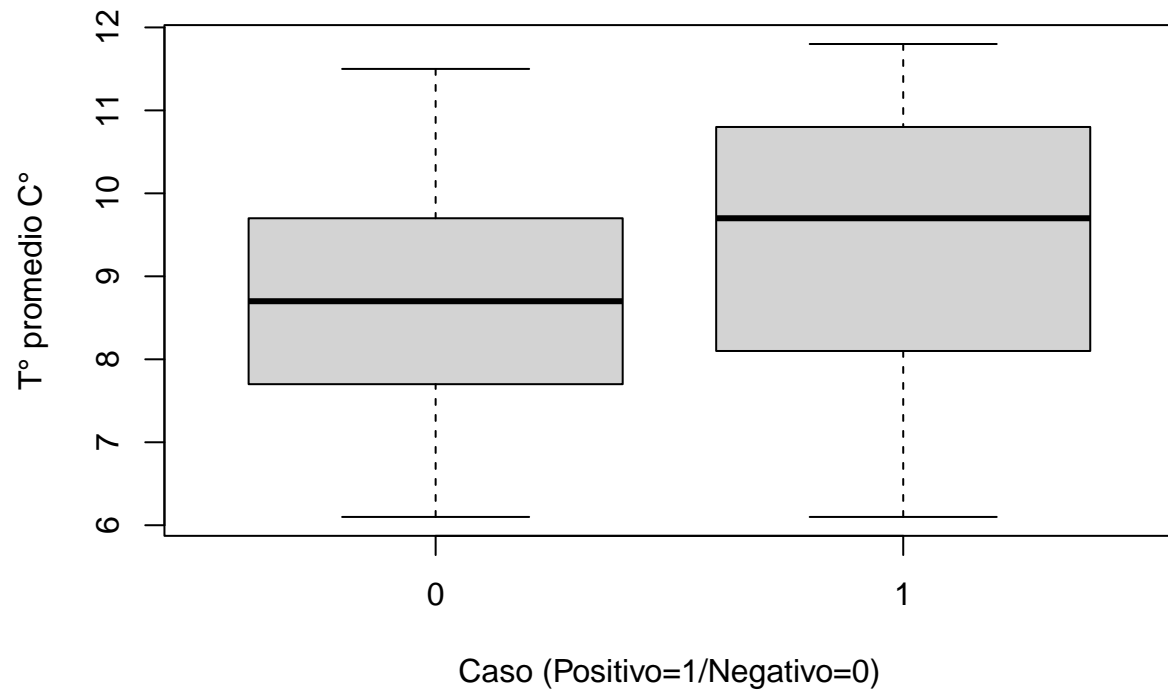


Gráfico Bloxpot asociando T° promedio con Caso

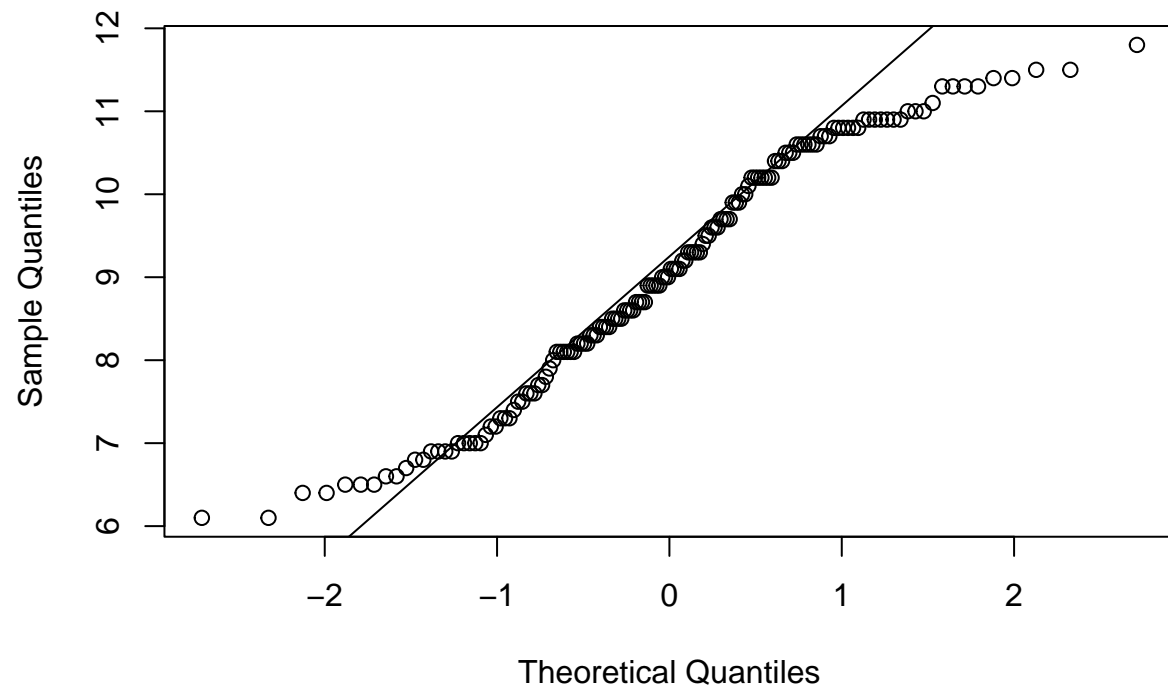
```
boxplot(D$MeanWaterTemperature~D$Case, xlab= "Caso (Positivo=1/Negativo=0)", ylab="T° promedio C°")
```



QQplot para ver la normalidad de los datos

```
qqnorm(D$MeanWaterTemperature)
qqline(D$MeanWaterTemperature)
```

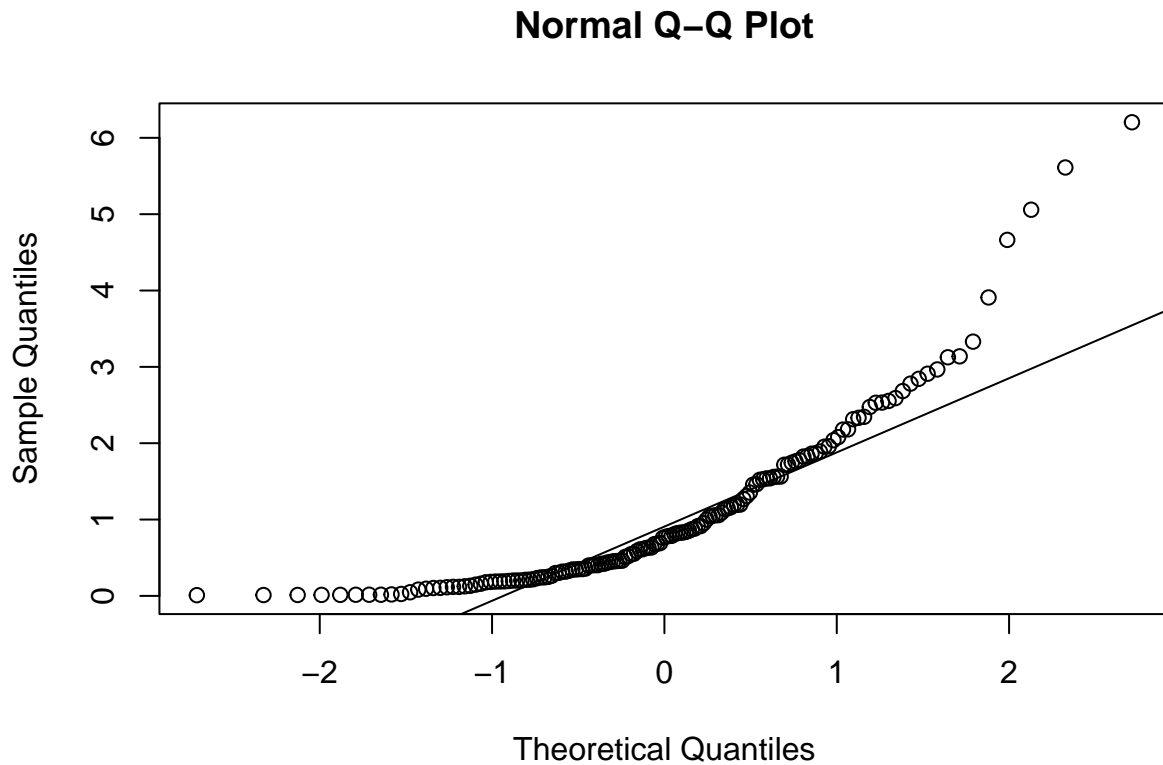
Normal Q-Q Plot



En este caso, buscaremos la normalidad con otra prueba, ya que la gran mayoría de las observaciones se

encuentran en el eje, por lo que la interpretación del valor de P, debería ayudarnos. Otra opción sería realizar un gráfico con una distribución gamma confirmando que los datos no tienen una distribución normal y una asimetría a la izquierda.

```
D$MeanWaterTemperature <- rgamma (150, 1)
qqnorm (D$MeanWaterTemperature)
qqline (D$MeanWaterTemperature)
```



Histograma Densidad (distribucion binomial con desplazamiento a la derecha)

```
ggplot(D, aes(x=D$Density))+
  geom_histogram(color="black", fill="grey", bins = 10)+theme_classic()+theme(text = element_text(size=
```

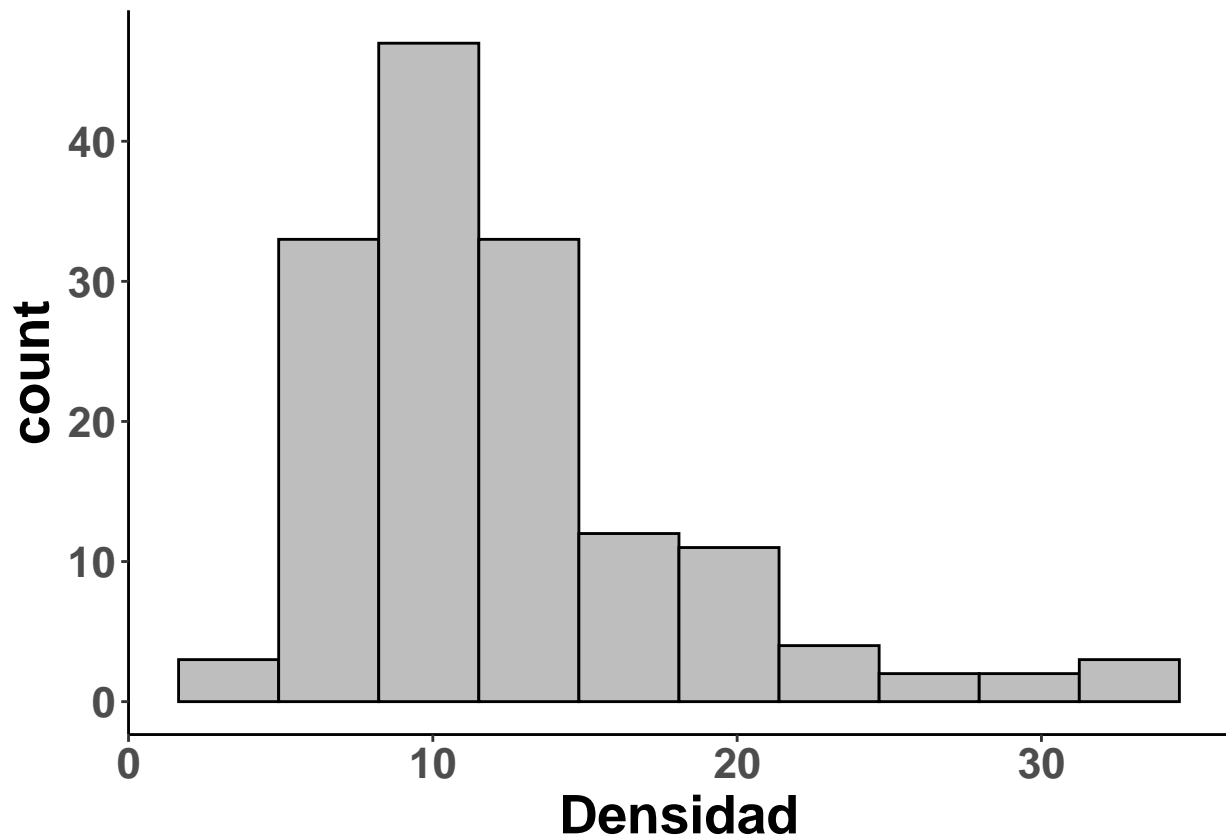
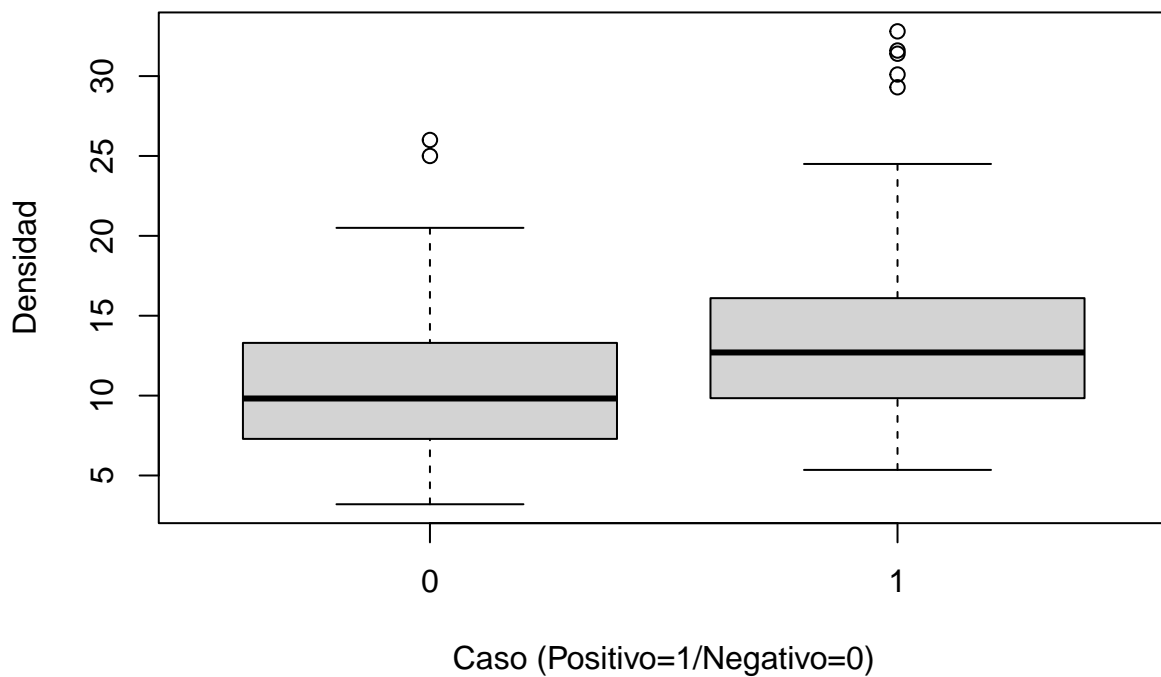


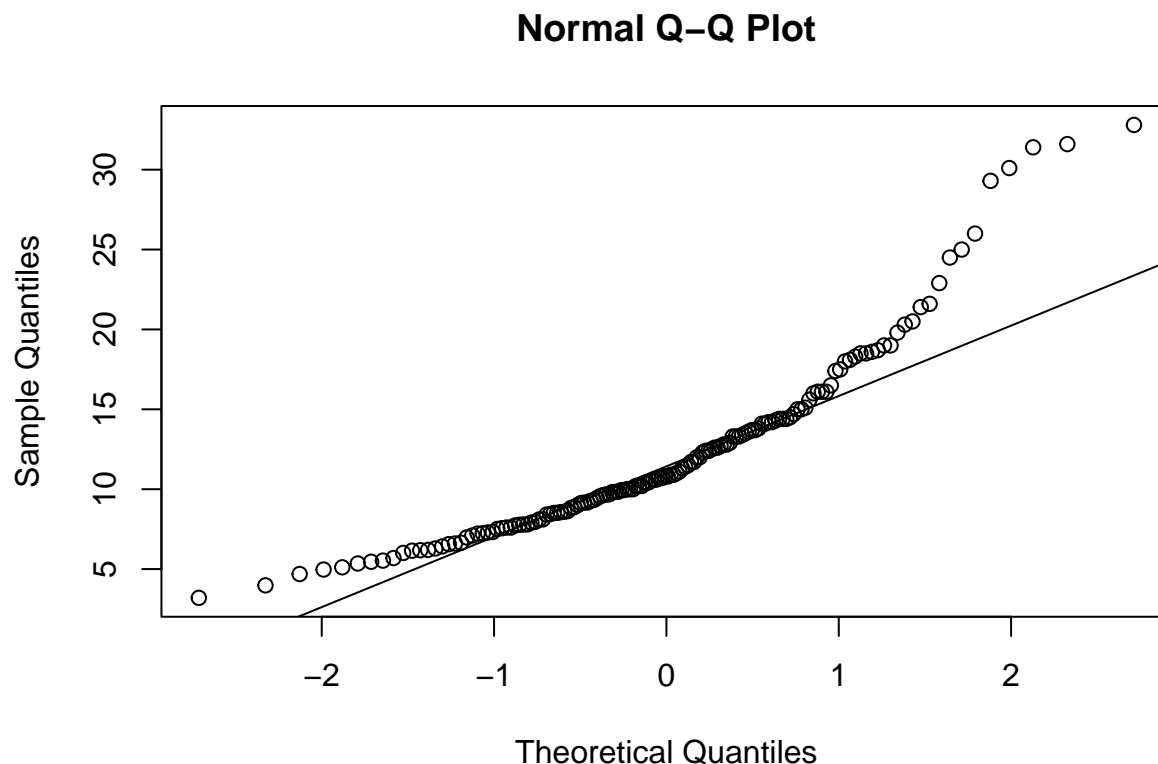
Gráfico Bloxpot asociando Densidad con Caso

```
boxplot(D$Density~D$Case, xlab= "Caso (Positivo=1/Negativo=0)", ylab="Densidad")
```



QQplot para ver la normalidad de los datos

```
qqnorm(D$Density)
qqline(D$Density)
```



De acuerdo a lo que vemos en el qqplot, los datos de Densidad no tienen una distribución normal y tienen una asimetría hacia la izquierda.

Prueba de Kolmogorov-Smirnov

```
ks.test(D$Density, "pnorm")
```

```
## Warning in ks.test.default(D$Density, "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

```
##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: D$Density
## D = 0.99931, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(D$Case, "pnorm")
```

```
## Warning in ks.test.default(D$Case, "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test
```



```
##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: D$Case
## D = 0.5, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Preguntas Adicionales: Hipotesis nula

¿La mortalidad observada en los casos positivos está explicada por tipo de vacuna, zona de manejo, temperatura promedio del agua, densidad de cultivo, compañía productora, enfermedad de branquias, presencia de caligus, calidad de smolt?

Comentario Final al analisis exploratorio de datos

La base de datos está limpia, completa, existen variables cualitativas y cuantitativas. Las variables cuantitativas no tienen una distribución normal por ende se requiere pruebas no paramétricas para un adecuado análisis estadístico.

Propuesta de Hipotesis

Hipotesis 0: La densidad no está asociada al caso (1) Hipotesis 1: La densidad está asociada al caso (1)

Para evaluar la correlación entre dos variables cuantitativas y no paramétricas se utilizará el coeficiente rho de Spearman.

```
cor.test(x=D$Density, y=D$Case, method='spearman')
```

```
## Warning in cor.test.default(x = D$Density, y = D$Case, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
## Spearman's rank correlation rho
##
## data: D$Density and D$Case
## S = 376818, p-value = 3.708e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.330072
```

Escala Spearman:

```
Correlación negativa perfecta..... -1
Correlación negativa fuerte moderada débil..... -0,5
Ninguna correlación..... 0
Correlación positiva moderada Fuerte..... +0,5
Correlación positiva perfecta..... + 1
```

La interpretación del coeficiente rho de Spearman concuerda en valores próximos a 1; indican una correlación fuerte y positiva. Valores próximos a -1 indican una correlación fuerte y negativa. Valores próximos a cero indican que no hay correlación lineal. Puede que exista otro tipo de correlación, pero no lineal. Y el P value es menor a 0,5 por lo tanto se acepta la hipótesis alternativa es decir hay correlación entre caso (1) y densidad.

Evaluación de supuestos

Independencia y análisis de residuales (falta graficar los supuestos) Se tiene en consideración que los supuestos se cumplen para los datos disponible porque son independientes y que no presentan error ya que en el análisis exploratorio previo se observó que existe igual numero de observaciones de densidad como números de jaulas que reportan mortalidad, estos datos, por ser colectados de un zona de fiordos con 5 areas de manejo (simil a un 'barrio' en Chile) donde se cultiva salmón del atlantico, donde se vieron afectados por eventos de mortalidad 87 de 150 centros de cultivo el año 202. Los datos de 150 centros fueron colectados mediante la misma metodología y de forma aleatoria proporcionando datos de densidad de cultivo (Kg/m³) acumulados en el tiempo, así como mortalidad caso 1 y 0, entendiendo según definición de Caso 1: son todas las jaulas que presentaron mortalidad igual o mayor a 5% día y Caso 0: todas aquellas jaulas cuya mortalidad fue menor a 5% día. Supuestos de tamaño de muestra En relación al tamaño de la muestra, se puede mencionar que corresponden al 100% de los datos observados.

#Normalidad En relación con la distribución de los datos de densidad, los cuales corresponden a variables cuantitativas continuas, estos no se ajustan a una distribución normal de acuerdo con el estadístico Prueba de Shapiro-wilks cuyo valor de p es de $>0,0001$. En relación con la mortalidad, estos datos corresponden a variables cuantitativa discreta y cuya distribución de tipo binomial a la derecha, no se ajusta a la mortalidad según Shapiro- wilks cuyo valor de p es de $>0,0001\%$. Si bien los datos cuya distribución no es normal, se estaría violando el supuesto, sin embargo, de acuerdo con la clase del Dr. Gallardo, se podría aceptados datos casi normales, particularmente si en n es $>a$ 30. La distribución a la derecha en el histograma puede estar explicado por el tiempo de cultivo de los centros en donde la densidad aumenta en la medida que aumenta el tiempo de cultivo o crecimiento de los peces, en este caso la media es de 12,34 kg/m³.

Conclusiones

Se acepta la hipótesis alternativa, la correlacion es positiva moderada, y muy posiblemente no lineal.