Proyecto Final de Estadística

Integrantes:
Laura Martir Beltrán C311
Adrián Hernández Castellanos C312
Yesenia Valdés Rodríguez C311

Dataset: SWISS

Medida de fertilidad estandarizada e indicadores socioeconómicos para cada una de las 47 provincias de habla francesa de Suiza alrededor de 1888. Contiene **47 observaciones** sobre **6 variables**, cada una de las cuales está en porcentaje, es decir, en [0,100].
Nuestras variables son:

- **Fertility**: Medida de fertilidad estandarizada común.
- **Agriculture**: % de hombres involucrados en la agricultura como ocupación.
- **Examination**: % de reclutas que obtienen la nota más alta en el examen militar.
- **Education**: % de educación más allá de la escuela primaria para los reclutas.
- **Catholic**: % 'Católico'(en oposición a 'protestante').
- **Infant.Mortality**: Nacidos vivos que viven menos de 1 año.

Todas las variables, excepto Fertility, dan proporciones de la población.

Fertility:

Medidas de Tendencia Central:

Media	Moda	Mediana
70.14255	68.06234	70.4

Con estos datos podemos apreciar que el nivel típico de fertilidad en las 47 provincias analizadas de Suiza es de un 70.14 %. La mitad de las provincias tiene una fertilidad por encima del 70.4 % y la otra mitad por debajo, lo que sugiere que los datos están relativamente equilibrados en torno a la media. Se puede apreciar que el valor de fertilidad del 68 % aproximadamente es el más frecuente entre las provincias, sin embargo, su diferencia respecto a la media y la mediana sugiere que puede haber una ligera asimetría en los datos.

Medidas de Dispersión:

Varianza	Desviación_Estandar	Mínimo	Máximo	Rango	Coeficiente_de_Variación
156.0425	12.4917	35	92.5	57.5	17.80901

Indica que existe una variabilidad considerable en los datos, con una desviación típica del 12.49 % en fertilidad respecto a la media. El coeficiente de variación con un valor del 17.81 % muestra que

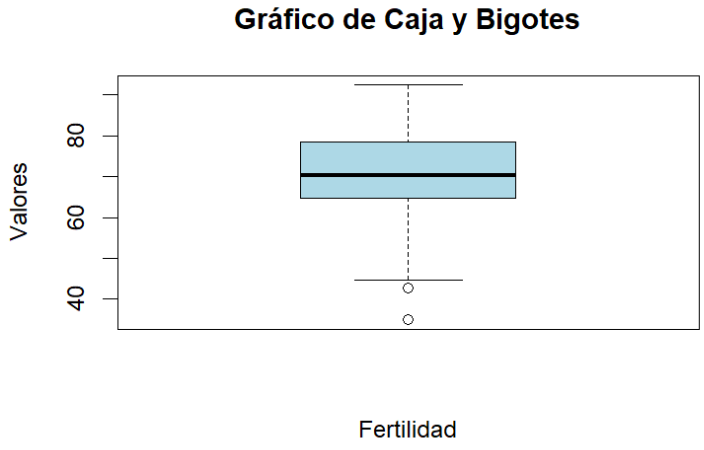
la dispersión relativa de la fertilidad es moderada. Esto indica que la fertilidad no está excesivamente dispersa respecto a la media, pero tampoco está completamente agrupada. La diferencia entre el orcentaje más alto (92.5 % en la provincia Franches-Mnt) y el más bajo (35 % en la provincia V. De Geneve) demuestra que hay una brecha amplia entre las provincias con mayor y menor fertilidad.
Conclusiones:

- Moderada Uniformidad: Aunque existe variabilidad en los niveles de fertilidad entre las provincias, la dispersión no es extrema.
- Distribución Centralizada: La similitud entre la media y la mediana sugiere una distribución centralizada de los datos, aunque podría haber valores atípicos que estén influyendo en la media.

Medidas de Posición:

```
  25%   50%   75%
64.70 70.40 78.45
```

Solo el 25 % de las provincias tiene valores significativamente bajos (<64.70 %) o altos (>78.45 %), lo que podría representar contextos socioeconómicos, culturales, o geográficos particulares. Estas medidas sugieren que la fertilidad en Suiza tiene un comportamiento centralizado, con pocos valores extremadamente altos o bajos.

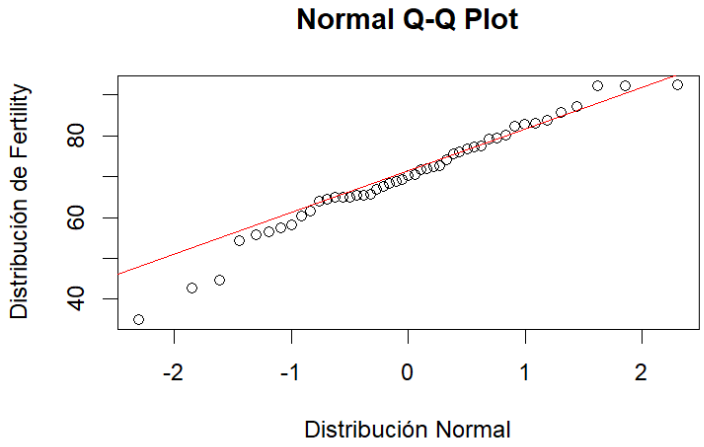


Distribución:

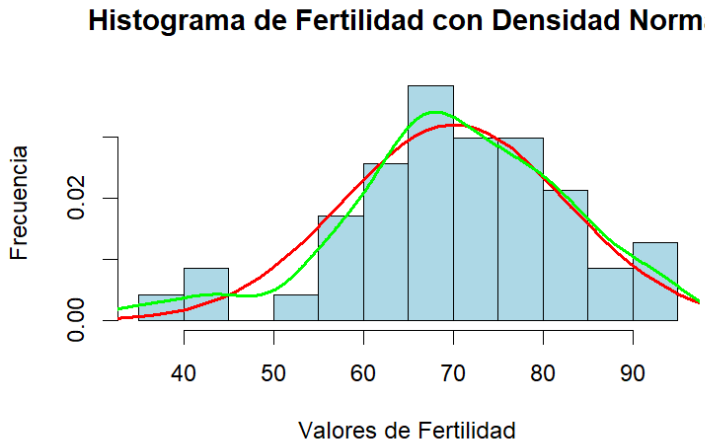
Podemos asegurar que la variable Fertility sigue una distribución normal?
Para aclarar esta duda, se procede a aplicar el test de Shapiro-Wilk que consiste en hacer prueba de hipótesis teniendo como hipótesis nula H_0 donde:

- H_0: la variable Fertlity sigue una distribución normal.
- El nivel de significancia es del 5 % ($\alpha = 0{,}05$).

El p-value resultante es 0.3449466 y al no ser menor que el nivel de significancia, no se rechaza la hipótesis nula.
En el caso de este dataset, es confiable usar Shapiro-Wilk dado que la muestra no es grande (n = 47) y en casos donde la muestra es grande, este test puede rechazar la hipotesis nula erróneamente.



Podemos ver que los valores de Fertility siguen una distribución normal para los valores más centralizados, presentando una ligera asimetría hacia la izquierda (sesgo negativo).



La fertilidad en las provincias suizas sigue un patrón general cercano a la normalidad, aunque hay cierta asimetría hacia los valores más bajos.

Análisis Teórico:

Intervalo de Confianza para la Media:

Debido a que este dataset no nos brinda información específica con respecto a la cantidad total de ciudadanos de las provincias y cuántos de estos cumplen deterinada propiedad, las medidas anteriormente presentadas son estimaciones de estas, en este caso, específicamente de la media real.
El intervalo de confianza para la media real de la variable Fertility al 95 % de confianza con varianza real desconocida es:

$$\left[\overline{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}; \overline{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] = \left[70{,}14255 - 2{,}012896\frac{12{,}4917}{\sqrt{47}}; 70{,}14255 + 2{,}012896\frac{12{,}4917}{\sqrt{47}}\right] =$$

[66,47485; 73,81025]

Intervalo de Confianza para la Varianza:

El intervalo de confianza de la varianza para Fertility al 95 % de confianza es:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}; \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right] = \left[\frac{(46)156{,}0425}{66{,}61653}; \frac{(46)156{,}0425}{29{,}16005}\right] = [107{,}7504; 246{,}1571]$$

Prueba de Hipótesis 1: Media Mayor o Igual que 66.5:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \leq 66{,}5$
- **Hipótesis Alternativa** (H_1): $\mu > 66{,}5$

Estadístico de Prueba:
Dado que la muestra es pequeña (n = 47) y la varianza poblacional es desconocida, usamos el estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{70{,}14255 - 66{,}5}{12{,}4917/\sqrt{47}} = 1{,}9991$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t > 1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media de la fertilidad en las provincias suizas es mayor o igual que 66.5 %.

Prueba de Hipótesis 2: Media Menor o Igual que 74:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \geq 74$
- **Hipótesis Alternativa** (H_1): $\mu < 74$

Estadístico de Prueba: Usamos el mismo estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{70{,}14255 - 74}{12{,}4917/\sqrt{47}} = -2{,}117$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t < -1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media de la fertilidad en las provincias suizas es menor o igual que 74 %.

Prueba de Hipótesis 1: Varianza Mayor o Igual que 107.75:

Hipótesis:

- **Hipótesis Nula** (H_0): $\sigma^2 \leq 107{,}75$
- **Hipótesis Alternativa** (H_1): $\sigma^2 > 107{,}75$

Estadístico de Prueba: Usamos el estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)156{,}0425}{107{,}75} = 66{,}6$$

Dado que el valor del estadístico de prueba cae en la región crítica($U > 62{,}8$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la varianza de la fertilidad en las provincias suizas es mayor o igual que 107,75.

Prueba de Hipótesis 2: Varianza Menor o Igual que 246:

Hipótesis:

- **Hipótesis Nula** (H_0): $\sigma^2 \geq 246$
- **Hipótesis Alternativa** (H_1): $\sigma^2 < 246$

Estadístico de Prueba Usamos el mismo estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)156{,}0425}{246} = 29{,}1786789$$

Dado que el valor del estadístico de prueba cae en la región crítica($U < 31{,}439$), **rechazamos la hipótesis nula** H_0. Por lo tanto, hay evidencia suficiente para concluir que la varianza de la fertilidad en las provincias suizas es menor o igual que 246.

Agriculture:

Medidas de Tendencia Central:

Media	Moda	Mediana
50.65957	62.65077	54.1

En promedio, aproximadamente el 50.66 % de la población en las provincias analizadas se dedica a la agricultura. Esto podría sugerir que la agricultura juega un papel importante en estas provincias, lo que podría influir en patrones de fertilidad(se verificará más adelante), ya que históricamente, en sociedades agrarias, las tasas de fertilidad tienden a ser más altas debido a factores como la necesidad de mano de obra. La mediana, ligeramente mayor que la media, indica que la mayoría de las provincias tienen más del 50 % de agricultores, esta sugiere una distribución ligeramente sesgada hacia la izquierda (debido a que la media es menor que la mediana y esta menor que la moda).

Medidas de Dispersión:

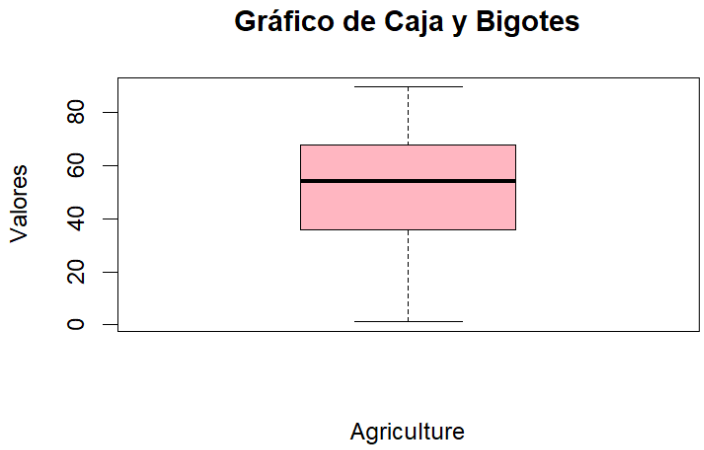
Varianza	Desviación_Estandar	Mínimo	Máximo	Rango	Coeficiente_de_Variación
515.7994	22.71122	1.2	89.7	88.5	44.83105

Los datos presentan una alta variabilidad, con una desviación estándar considerablemente grande en relación con la media. Esto indica que el porcentaje de agricultores varía mucho entre provincias. Existe una gran diferencia entre la provincia con el menor porcentaje de agricultores (1.2 % en V. De Geneve) y la mayor (89.7 % en Herens), lo que refuerza la heterogeneidad de las provincias.

Medidas de Posición:

25%	50%	75%
35.90	54.10	67.65

Las provincias con porcentajes inferiores al primer cuartil (<35.90 %) probablemente tengan mayor diversificación económica o urbanización, mientras que aquellas superiores al tercer cuartil (>67.65 %) podrían tener economías más centradas en la agricultura.

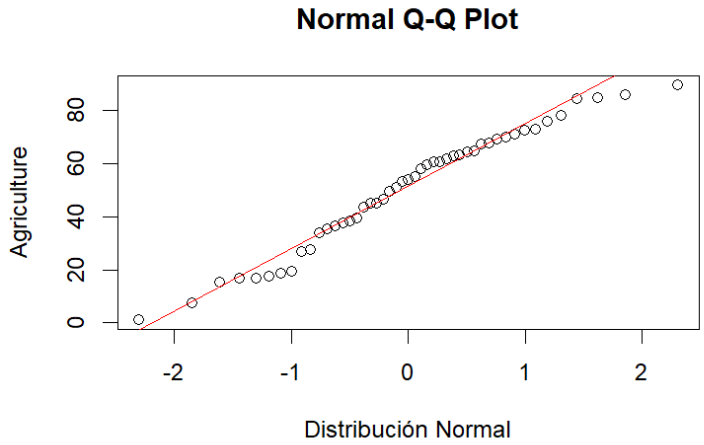


Distribución:

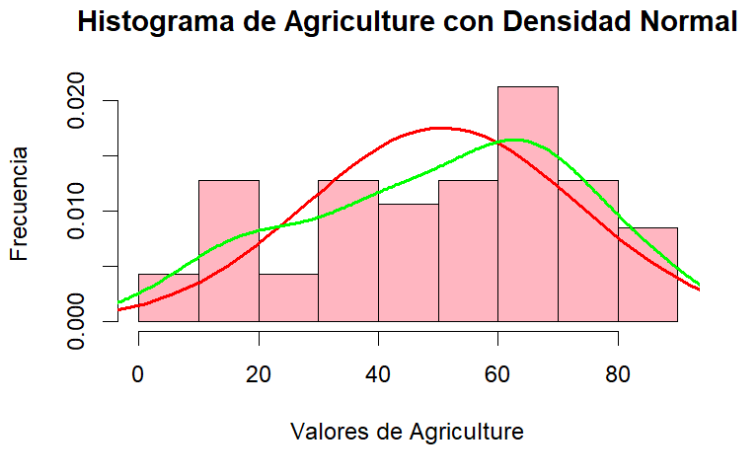
Podemos asegurar que la variable Agriculture sigue una distribución normal?
Aplicar el test de Shapiro-Wilk:

- H_0: la variable Agriculture sigue una distribución normal.
- El nivel de significancia es del 5 % ($\alpha = 0{,}05$).

El p-value resultante es 0.1930223 y al no ser menor que el nivel de significancia, no se rechaza la hipótesis nula.



Aquí ocurre lo mismo que con la variable Fertility, lo que con una variación mucho mayor. Distribución normal para datos más centralizados y asimetría hacia la derecha(sesgo negativo).



A pesar de que no podemos rechazar la hipótesis nula de que Agriculture siga una distribución normal, tampoco la podemos aceptar debido a la alta variabilidad que tiene sus valores provocando una asimetría que no se puede ignorar.

Análisis Teórico:

Intervalo de Confianza para la Media:

Elegimos un nivel de confianza del 95 %, por lo que $\alpha = 0{,}05$.
El intervalo de confianza para la media de la variable Agriculture al 95 % de confianza donde la varianza es desconocida es:

$$\left[\overline{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}; \overline{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] = \left[50{,}65957 - 2{,}012896\frac{22{,}71122}{\sqrt{47}}; 50{,}65957 + 2{,}012896\frac{22{,}71122}{\sqrt{47}}\right] =$$
$$[43{,}99131; 57{,}32784]$$

Intervalo de Confianza para la Varianza:

El intervalo de confianza de la varianza para Agriculture al 95 % de confianza es:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}; \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right] = \left[\frac{(46)515{,}7994}{66{,}61653}; \frac{(46)515{,}7994}{29{,}16005}\right] = [356{,}1695; 813{,}6738]$$

Prueba de Hipótesis 1: Media Mayor o Igual que 44:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \leq 44$
- **Hipótesis Alternativa** (H_1): $\mu > 44$

Estadístico de Prueba:
Dado que la muestra es pequeña (n = 47) y la varianza poblacional es desconocida, usamos el estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{50{,}65957 - 44}{22{,}71122/\sqrt{47}} = 2{,}0103$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t > 1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media del porcentaje de hombres involucrados en la agricultura en las provincias suizas es mayor o igual que 44 %.

Prueba de Hipótesis 2: Media Menor o Igual que 57.3:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \geq 57{,}3$
- **Hipótesis Alternativa** (H_1): $\mu < 57{,}3$

Estadístico de Prueba: Usamos el mismo estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{50{,}65957 - 57{,}3}{22{,}71122/\sqrt{47}} = -2{,}0045$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t < -1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media del porcentaje de hombres involucrados en la agricultura en las provincias suizas es menor o igual que 57.3 %.

Prueba de Hipótesis 1: Varianza Mayor o Igual que 356.17:

Hipótesis:

- **Hipótesis Nula** (H_0): $\sigma^2 \leq 356{,}17$
- **Hipótesis Alternativa** (H_1): $\sigma^2 > 356{,}17$

Estadístico de Prueba: Usamos el estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)515{,}7994}{356{,}17} = 66{,}6$$

Dado que el valor del estadístico de prueba cae en la región crítica($U > 62{,}8$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la varianza del porcentaje de hombres involucrados en la agricultura en las provincias suizas es mayor o igual que 356,17.

Prueba de Hipótesis 2: Varianza Menor o Igual que 813.67:

Hipótesis:

- **Hipótesis Nula** (H_0): $\sigma^2 \geq 813{,}67$
- **Hipótesis Alternativa** (H_1): $\sigma^2 < 813{,}67$

Estadístico de Prueba Usamos el mismo estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)515{,}7994}{813{,}67} = 29,1601907$$

Dado que el valor del estadístico de prueba cae en la región crítica($U < 31{,}439$), **rechazamos la hipótesis nula** H_0. Por lo tanto, hay evidencia suficiente para concluir que la varianza del porcentaje de hombres involucrados en la agricultura en las provincias suizas es menor o igual que 813,67.

Examination:

Medidas de Tendencia Central:

Media	Moda	Mediana
16.48936	14	16

En promedio, los reclutas obtuvieron una puntuación relativamente baja en comparación con el rango total (3 % a 37 %). La media y la mediana son cercanas, lo que sugiere una distribución relativamente simétrica, sin embargo, la moda más baja indica que hay un número considerable de reclutas con puntuaciones por debajo del promedio, lo cual podría ser motivo de preocupación para el proceso de selección militar.

Medidas de Dispersión:

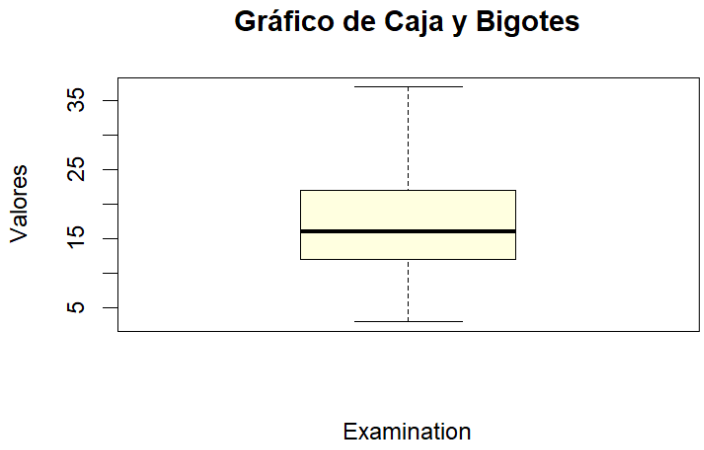
Varianza	Desviación_Estandar	Mínimo	Máximo	Rango	Coeficiente_de_Variación
63.64662	7.977883	3	37	34	48.382

Existe una dispersión considerable en las puntuaciones, esto sugiere que hay una variabilidad notable en el rendimiento de los reclutas. Existe más de una provincia con el menor porcentaje de reclutas con buenas califcaciones(3 %), estas son Sierre y Conthey. Luego, solo V. De Geneve tiene el mayor porcentaje el cual sigue siendo bastante bajo(37 %). Se puede ver una gran diferencia en preparación y habilidades entre las provincias.

Medidas de Posición:

25%	50%	75%
12	16	22

El 75 % de las 47 provincias analizadas tienen un porcentaje de reclutas con buenas notas menor que 22 %(bastante bajo).

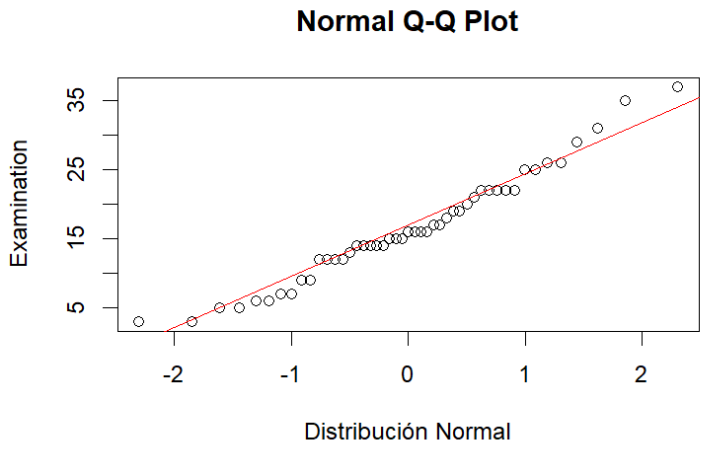


Distribución:

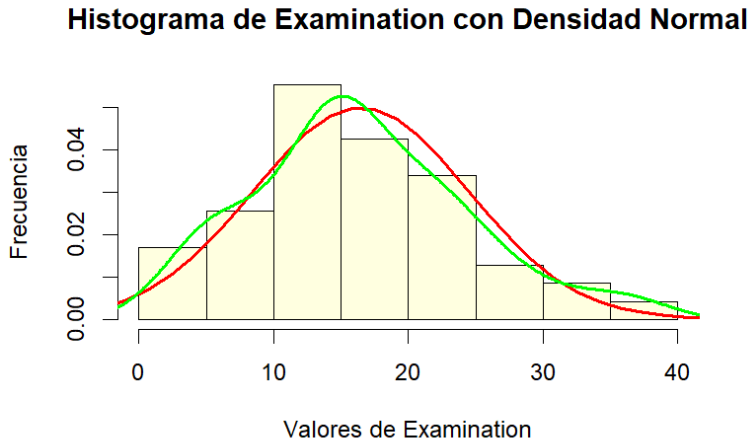
Podemos asegurar que la variable Examination sigue una distribución normal?
Aplicar el test de Shapiro-Wilk:

- H_0: la variable Examination sigue una distribución normal.
- El nivel de significancia es del 5 % ($\alpha = 0{,}05$).

El p-value resultante es 0.2562701 y al no ser menor que el nivel de significancia, curiosamente no se rechaza la hipótesis nula.



Podemos apreciar de que sorprendentemente los datos no se alejan tanto de la distribución normal. Presentando una ligera asimetría hacia la izquierda (sesgo positivo) debido al bajo porciento de reclutas con buena nota en las 47 provincias.



Análisis Teórico:

Intervalo de Confianza para la Media:

Elegimos un nivel de confianza del 95 %, por lo que $\alpha = 0{,}05$.
El intervalo de confianza para la media de la variable Agriculture al 95 % de confianza donde la

varianza es desconocida es:

$$\left[\overline{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}; \overline{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] = \left[16{,}48936 - 2{,}012896\frac{7{,}977883}{\sqrt{47}}; 16{,}48936 + 2{,}012896\frac{7{,}977883}{\sqrt{47}}\right] =$$

$$[14{,}14697; 18{,}83176]$$

Intervalo de Confianza para la Varianza:

El intervalo de confianza de la varianza para Agriculture al 95 % de confianza es:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}; \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right] = \left[\frac{(46)63{,}64662}{66{,}61653}; \frac{(46)63{,}64662}{29{,}16005}\right] = [43{,}94922; 100{,}40258]$$

Prueba de Hipótesis 1: Media Mayor o Igual que 14:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \leq 14$
- **Hipótesis Alternativa** (H_1): $\mu > 14$

Estadístico de Prueba:
Dado que la muestra es pequeña (n = 47) y la varianza poblacional es desconocida, usamos el estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{16{,}48936 - 14}{7{,}977883/\sqrt{47}} = 2{,}1392$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t > 1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media del porcentaje de reclutas que obtienen la nota más alta en el examen militar en las provincias suizas es mayor o igual que 14 %.

Prueba de Hipótesis 2: Media Menor o Igual que 18.8:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \geq 18{,}8$
- **Hipótesis Alternativa** (H_1): $\mu < 18{,}8$

Estadístico de Prueba: Usamos el mismo estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{16{,}48936 - 18{,}8}{7{,}977883/\sqrt{47}} = -1{,}9856$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t < -1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media del porcentaje de reclutas que obtienen la nota más alta en el examen militar en las provincias suizas es menor o igual que 18.8 %.

Prueba de Hipótesis 1: Varianza Mayor o Igual que 43.9:

Hipótesis:

- **Hipótesis Nula** (H_0): $\sigma^2 \leq 43{,}9$
- **Hipótesis Alternativa** (H_1): $\sigma^2 > 43{,}9$

Estadístico de Prueba: Usamos el estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)63{,}64662}{43{,}9} = 66{,}6912191$$

Dado que el valor del estadístico de prueba cae en la región crítica($U > 62{,}8$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la varianza del porcentaje de reclutas que obtienen la nota más alta en el examen militar en las provincias suizas es mayor o igual que 43,9.

Prueba de Hipótesis 2: Varianza Menor o Igual que 100.4:

Hipótesis:

- **Hipótesis Nula** (H_0): $\sigma^2 \geq 100{,}4$
- **Hipótesis Alternativa** (H_1): $\sigma^2 < 100{,}4$

Estadístico de Prueba Usamos el mismo estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)63{,}64662}{100{,}4} = 29{,}160802$$

Dado que el valor del estadístico de prueba cae en la región crítica($U < 31{,}439$), **rechazamos la hipótesis nula** H_0. Por lo tanto, hay evidencia suficiente para concluir que la varianza del porcentaje de reclutas que obtienen la nota más alta en el examen militar en las provincias suizas es menor o igual que 100,4.

Infant.Mortality:

Medidas de Tendencia Central:

Media	Moda	Mediana
19.94255	18	20

La moda (19.91 %) y la mediana (20 %) son valores muy cercanos a la media(19.94 %), lo que sugiere que la distribución de las tasas de mortalidad es relativamente simétrica y no presenta grandes asimetrías.

Medidas de Dispersión:

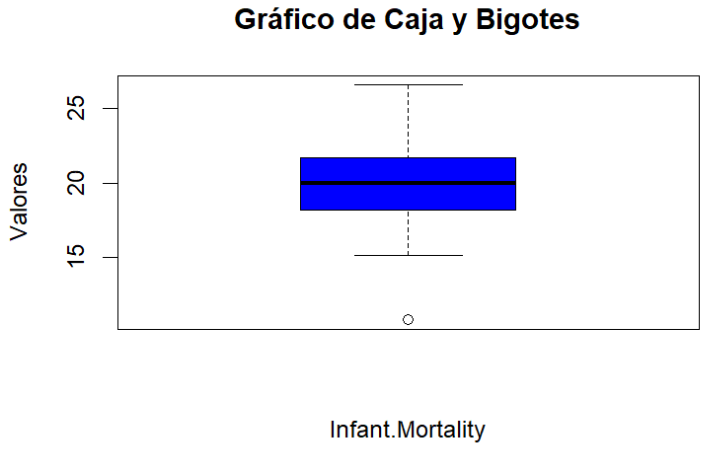
Varianza	Desviación_Estandar	Mínimo	Máximo	Rango	Coeficiente_de_Variación
8.483802	2.912697	10.8	26.6	15.8	14.60544

Hay una variabilidad moderada en las tasas de mortalidad infantil entre las provincias. Una desviación estándar de 2.91 % sugiere que la mayoría de las provincias tienen tasas de mortalidad infantil que se encuentran dentro de un rango razonable alrededor de la media, sugiriendo que las diferencias entre provincias no son excesivas. Estos datos sugieren que Suiza tiene un nivel relativamente bajo de mortalidad infantil, aunque existen diferencias significativas entre provincias que podrían requerir atención específica para abordar las causas subyacentes en aquellas con tasas más altas.

Medidas de Posición:

25%	50%	75%
18.15	20.00	21.70

La distancia entre el primer cuartil (18.15 %) y el tercer cuartil (21.70 %) es de 3.55 %, lo que indica una dispersión moderada en las tasas de mortalidad infantil entre las provincias. La mediana (20.00 %) está en el medio del rango, lo que sugiere que no hay una gran asimetría en los datos.

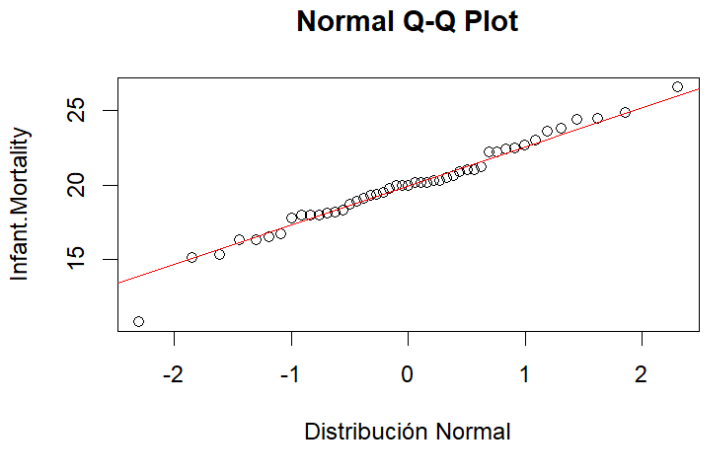


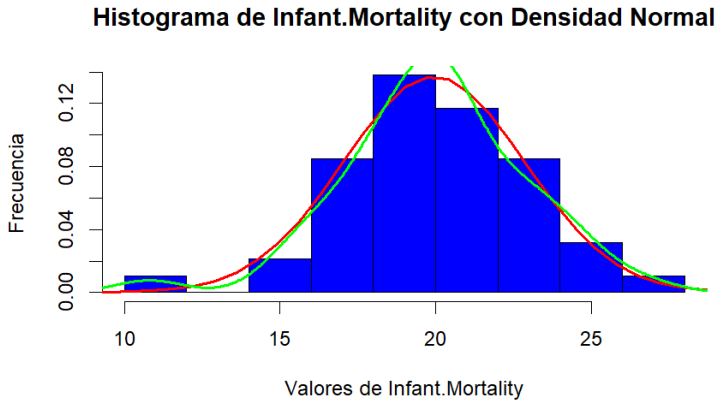
Distribución:

Podemos asegurar que la variable Infant.Mortality sigue una distribución normal?
Aplicar el test de Shapiro-Wilk:

- H_0: la variable Infant.Mortality sigue una distribución normal.
- El nivel de significancia es del 5 % ($\alpha = 0{,}05$).

El p-value resultante es 0.4978056 y al no ser menor que el nivel de significancia, no se rechaza la hipótesis nula.





Análisis Teórico:

Intervalo de Confianza para la Media:

Elegimos un nivel de confianza del 95 %, por lo que $\alpha = 0{,}05$.
El intervalo de confianza para la media de la variable INfant Mortality al 95 % de confianza donde la varianza es desconocida es:

$$\left[\overline{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}; \overline{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] = \left[19{,}94255 - 2{,}012896\frac{2{,}912697}{\sqrt{47}}; 19{,}94255 + 2{,}012896\frac{2{,}912697}{\sqrt{47}}\right] =$$

$$[19{,}08735; 20{,}79775]$$

Intervalo de Confianza para la Varianza:

El intervalo de confianza de la varianza para Agriculture al 95 % de confianza es:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}; \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right] = \left[\frac{(46)8{,}483802}{66{,}61653}; \frac{(46)8{,}483802}{29{,}16005}\right] = [5{,}858229; 13{,}383202]$$

Prueba de Hipótesis 1: Media Mayor o Igual que 19:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \leq 19$

- **Hipótesis Alternativa** (H_1): $\mu > 19$

Estadístico de Prueba:
Dado que la muestra es pequeña (n = 47) y la varianza poblacional es desconocida, usamos el estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{19{,}94255 - 19}{2{,}912697/\sqrt{47}} = 2{,}2185$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t > 1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media de la tasa de mortalidad infantil en las provincias suizas es mayor o igual que 19 %.

Prueba de Hipótesis 2: Media Menor o Igual que 20.8:

Hipótesis:

- **Hipótesis Nula** (H_0): $\mu \geq 20{,}8$
- **Hipótesis Alternativa** (H_1): $\mu < 20{,}8$

Estadístico de Prueba: Usamos el mismo estadístico t de Student:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{19{,}94255 - 20{,}8}{2{,}912697/\sqrt{47}} = -2{,}0182$$

Dado que el valor del estadístico de prueba cae en la región crítica ($t < -1{,}67866$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la media de la tasa de mortalidad infantil en las provincias suizas es menor o igual que 20.8 %.

Prueba de Hipótesis 1: Varianza Mayor o Igual que 5.85:

5.858229; 13.383202 **Hipótesis:**

- **Hipótesis Nula** (H_0): $\sigma^2 \leq 5{,}85$
- **Hipótesis Alternativa** (H_1): $\sigma^2 > 5{,}85$

Estadístico de Prueba: Usamos el estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)8{,}483802}{5{,}85} = 66{,}7102379$$

Dado que el valor del estadístico de prueba cae en la región crítica($U > 62{,}8$), **rechazamos la hipótesis nula** H_0 a favor de la hipótesis alternativa H_1. Por lo tanto, hay evidencia suficiente para concluir que la varianza de la tasa de mortalidad infantil en las provincias suizas es mayor o igual que 5,85.

Prueba de Hipótesis 2: Varianza Menor o Igual que 13.4:

Hipótesis:

- **Hipótesis Nula** (H_0): $\sigma^2 \geq 13{,}4$
- **Hipótesis Alternativa** (H_1): $\sigma^2 < 13{,}4$

Estadístico de Prueba Usamos el mismo estadístico U para la varianza:

$$U = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(46)8{,}483802}{13{,}4} = 29{,}1234994$$

Dado que el valor del estadístico de prueba cae en la región crítica($U < 31{,}439$), **rechazamos la hipótesis nula** H_0. Por lo tanto, hay evidencia suficiente para concluir que la varianza de la tasa de mortalidad infantil en las provincias suizas es menor o igual que 13,4.

Education:

Medidas de Tendencia Central:

Media	Moda	Mediana
10.97872	7	8

En promedio, solo un 10.98 % de los reclutas tiene educación más allá de la escuela primaria. Este valor es relativamente bajo, sugiriendo que la mayoría de los reclutas no han alcanzado niveles educativos superiores. El 50 % de las provincias evaluadas tiene un porcentaje menor del 8 % de reclutas con un nivel de educación superior al básico, esto refuerza la idea de que la mayoría tiene una educación limitada, lo cual podría ser un punto crítico para el desarrollo del capital humano en el contexto militar.

Medidas de Dispersión:

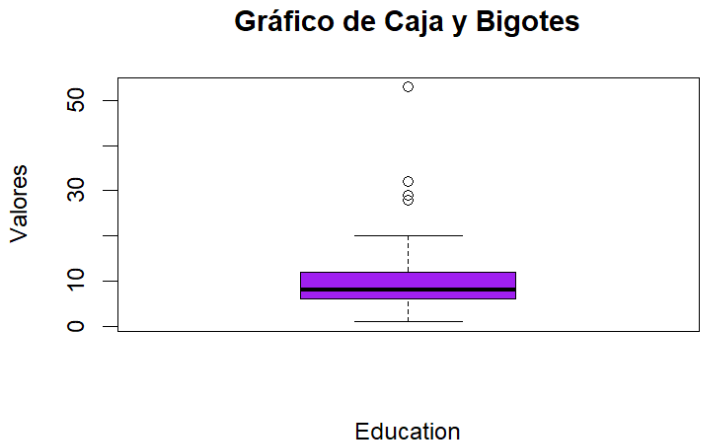
Varianza	Desviación_Estandar	Mínimo	Máximo	Rango	Coeficiente_de_Variación
92.45606	9.615407	1	53	52	87.5822

Alta variabilidad en los porcentajes de educación entre las provincias. Oron, la provincia con menor porcentaje de reclutas con mayor nivel educativo (1 %) y V. De Geneve de las únicas provincia con un porcentaje mayor al promedio (53 %). La alta varianza y el rango indican que hay diferencias notables entre provincias en cuanto a la educación superior de los reclutas. Algunas provincias pueden estar proporcionando mejores oportunidades educativas que otras.

Medidas de Posición:

25%	50%	75%
6	8	12

El tercer cuartil (75 %) indica que el 75 % de las provincias evaluadas tiene un porcentaje de reclutas con educación superior a la básica igual o inferior a 12 %. Esto implica que solo el 25 % restante tiene porcentajes superiores a este valor, lo que sugiere que hay un grupo pequeño que logra niveles educativos más altos.

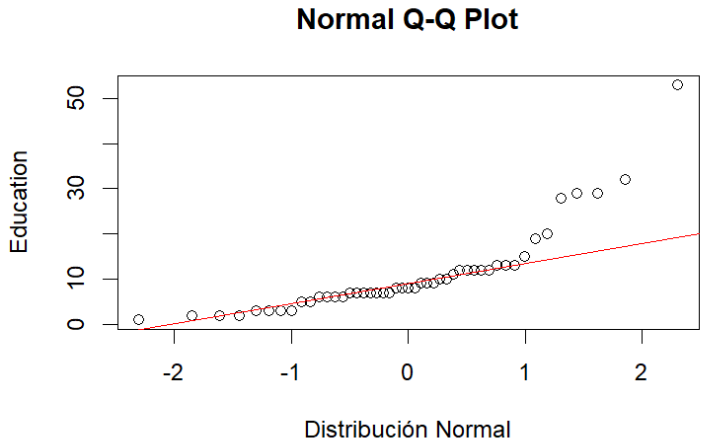


Distribución:

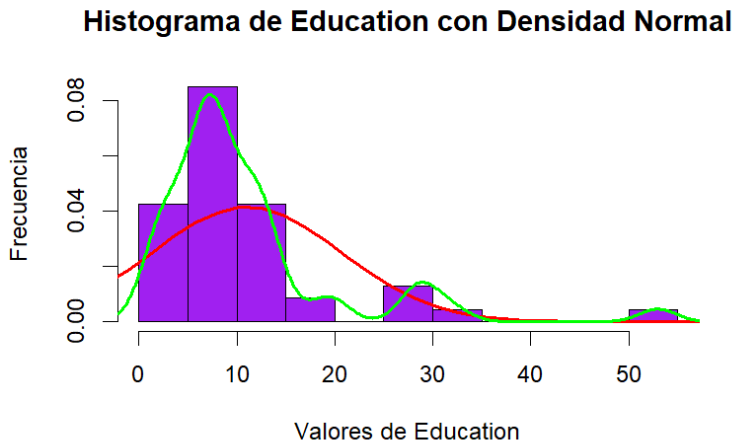
Dadas las características anteriormente descritas, se puede sospechar que esta variable no sigue una distribución normal, confirmemos esta hipótesis.
Aplicar el test de Shapiro-Wilk:

- H_0: la variable Education sigue una distribución normal.
- El nivel de significancia es del 5 % ($\alpha = 0{,}05$).

El p-value resultante es 1.31202e-07 y claramente podemos ver que es muy inferior al nivel de significancia, por lo tanto podemos rechazar la hipótesis nula y concluir de que la variable Education no sigue una distribución normal.



Podemos ver que es asimétrica hacia la izquierda.



Análisis Teórico:

Dado que la variable Education no distribuye normal, no podemos hallar los intervalos de confianza de la media y varianza con los métodos aplicados anteriormente. En este caso usaremos un método llamado BCa(Bias-Corrected and Accelerated) Bootstrap el cual es una versión mejorada del método Bootstrap que corrige el sesgo y la aceleración en la estimación de los intervalos de confianza, particularmente útil para datos no normales y asimétricos como este, ya que proporciona intervalos de confianza más precisos y ajustados.

Intervalo de Confianza para la Media:

El intervalo de confianza para la media real de la variable Education al 95 % de confianza con varianza real desconocida es:

$$[8{,}81; 14{,}72]$$

Intervalo de Confianza para la Varianza:

El intervalo de confianza de la varianza para Education al 95 % de confianza es:

$$[44{,}09; 230{,}80]$$

Catholic:

Medidas de Tendencia Central:

Media	Moda	Mediana
41.14383	8.159039	15.14

El porcentaje más frecuente de católicos en las provincias es bastante bajo, lo que implica que muchas provincias tienen un porcentaje de católicos muy por debajo del promedio. La mitad de las provincias tiene un porcentaje de católicos inferior al 15.14 %.

Medidas de Dispersión:

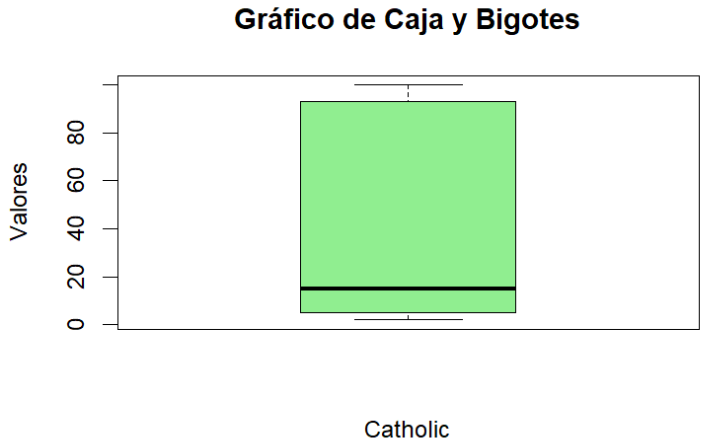
Varianza	Desviación_Estandar	Mínimo	Máximo	Rango	Coeficiente_de_Variación
1739.295	41.70485	2.15	100	97.85	101.3636

Alta dispersión en los porcentajes de católicos entre las provincias. Esto sugiere que algunas provincias tienen un porcentaje muy alto de católicos, mientras que otras tienen porcentajes muy bajos. Tenemos la provincia La Vallee con un porcentaje del 2.15 % y Herens donde el 100 % de su población es católica. Algunas provincias tienen una población católica considerablemente más alta, mientras que otras tienen una representación muy baja, lo que podría reflejar diferencias culturales, históricas o socioeconómicas entre las regiones.

Medidas de Posición:

25%	50%	75%
5.195	15.140	93.125

Las medidas de posición indican una clara desigualdad en la representación católica entre las provincias suizas. Con un primer cuartil bajo (5.195 %), se evidencia que una parte significativa del grupo tiene poca o ninguna representación católica (la mayoría son protestantes), mientras que el tercer cuartil alto (93.125 %) sugiere que hay provincias donde la religión sigue siendo muy relevante.

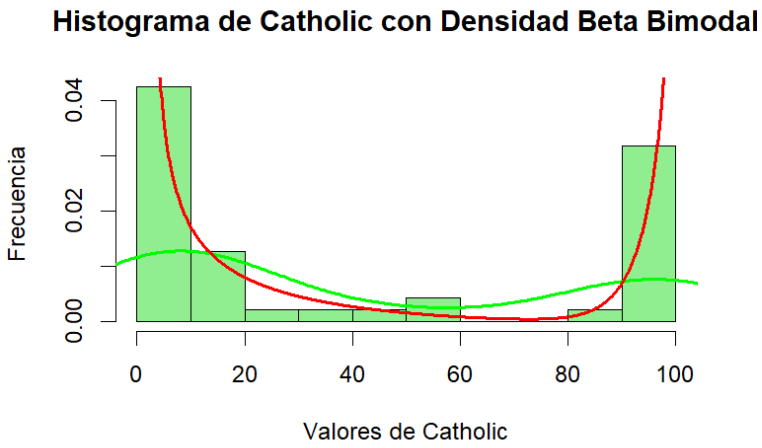


Distribución:

Dadas las características anteriormente descritas, se puede sospechar que esta variable no sigue una distribución normal, confirmemos esta hipótesis.
Aplicar el test de Shapiro-Wilk:

- H_0: la variable Catholic sigue una distribución normal.
- El nivel de significancia es del 5 % ($\alpha = 0{,}05$).

El p-value resultante es 1.20461e-07 y claramente podemos ver que es muy inferior al nivel de significancia, por lo tanto podemos rechazar la hipótesis nula y concluir de que la variable Catholic no sigue una distribución normal.



Análisis Teórico:

Dado que la variable Catholic no distribuye normal, no podemos hallar los intervalos de confianza de la media y varianza con los métodos aplicados anteriormente. En este caso usaremos un método llamado BCa(Bias-Corrected and Accelerated) Bootstrap el cual es una versión mejorada del método Bootstrap que corrige el sesgo y la aceleración en la estimación de los intervalos de confianza, particularmente útil para datos no normales y asimétricos como este, ya que proporciona intervalos de confianza más precisos y ajustados.

Intervalo de Confianza para la Media:

El intervalo de confianza para la media real de la variable Catholic al 95 % de confianza con varianza real desconocida es:

$$[29{,}88; 55{,}19]$$

Intervalo de Confianza para la Varianza:

El intervalo de confianza de la varianza para Catholic al 95 % de confianza es:

$$[1394; 2000]$$

Categorizando las variables:

Ninguna de las variables de nuestro dataset es categórica por lo tanto, para poder realizar análisis estadísticos que dependan de estas es necesario categorizarlas.

Fertility:

Clasificaremos la fertilidad de las 47 provincias en tres categorías:

- **Baja**: Cuando el porcentaje pertenece al intervalo [35, 65.43333].
- **Media**: Cuando el porcentaje pertenece al intervalo (65.43333, 75.9].
- **Alta**: Cuando el porcentaje pertenece al intervalo (75.9, 92.5].

Estos intervalos están definidos con respecto a los terciles de Fertility tal que los tres tienen la misma probabilidad.

Agriculture:

Clasificaremos las 47 provincias de Suiza analizadas en base al porcentaje de agricultores en cada una de ellas:

- **Urbanitos**: Cuando el porcentaje pertenece al intervalo [1.2, 40.96667].
- **Agr_Urb**: Cuando el porcentaje pertenece al intervalo (40.96667, 63.36667].
- **Agricultores**: Cuando el porcentaje pertenece al intervalo (63.36667, 89.7].

Estos intervalos están definidos con respecto a los terciles de Agriculture tal que los tres tienen la misma probabilidad.

Examination:

Clasificaremos los resultados de los exámenes en tres categorías:

- **Peor**: Cuando el porcentaje pertenece al intervalo [3, 14].
- **Normal**: Cuando el porcentaje pertenece al intervalo (14, 19].
- **Mejor**: Cuando el porcentaje pertenece al intervalo (19, 37].

Estos intervalos están definidos con respecto a los terciles de Examination, tal que los tres tiene la misma probabilidad.

Education:

Clasificaremos el nivel de educación de las provincias en tres categorías:

- **Mala**: Cuando el porcentaje pertenece al intervalo [1, 7].
- **Normal**: Cuando el porcentaje pertenece al intervalo (7, 11.66667].
- **Buena**: Cuando el porcentaje pertenece al intervalo (11.66667, 53].

Estos intervalos están definidos con respecto a los terciles de Education, tal que los tres tiene la misma probabilidad.

Catholic:

Clasificaremos la religión de las provincias en tres categorías:

- **Protestantes**: Cuando el porcentaje pertenece al intervalo [2.15, 6.64].
- **Mezclados**: Cuando el porcentaje pertenece al intervalo (6.64, 76.00333].
- **Católicos**: Cuando el porcentaje pertenece al intervalo (76.00333, 100].

Estos intervalos están definidos con respecto a los terciles de Catholic, tal que los tres tiene la misma probabilidad.

Infant.Mortality:

Clasificaremos la mortalidad infantil de las provincias en tres categorías:

- **Baja**: Cuando el porcentaje pertenece al intervalo [10.8, 18.96667].
- **Media**: Cuando el porcentaje pertenece al intervalo (18.96667, 20.8].
- **Alta**: Cuando el porcentaje pertenece al intervalo (20.8, 26.6].

Estos intervalos están definidos con respecto a los terciles de Infant.Mortality, tal que los tres tiene la misma probabilidad.

Tablas de Contingencia:

En base a las categorizaciones de nuestras variables anteriormente descritas, se porcede a crear las tablas de contingencia para cada una.

Fertility-Agriculture:

	Urbanitos	Agr_Urb	Agricultores	Sum
Baja	6	7	3	16
Media	5	3	7	15
Alta	5	5	6	16
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde:

- r: Número de filas de la tabla de contingencia.
- c: Número de columnas de la tabla de contingencia.
- O_{ij}: Frecuencia observada en la celda (i, j).
- E_{ij}: Frecuencia esperada en la celda (i, j), calculada como:

$$E_{ij} = \frac{\text{Total fila } i \times \text{Total columna } j}{\text{Total general}}$$

$$E_{11} = E_{13} = E_{31} = E_{33} = \frac{16 \times 16}{47} = 5{,}446809$$

$$E_{12} = E_{21} = E_{23} = E_{32} = \frac{15 \times 16}{47} = 5{,}106383$$

$$E_{22} = \frac{15 \times 15}{47} = 4{,}787234$$

$$\chi^2 = \sum_{j=1}^{c} \frac{(O_{1j} - E_{1j})^2}{E_{1j}} + \sum_{j=1}^{c} \frac{(O_{2j} - E_{2j})^2}{E_{2j}} + \sum_{j=1}^{c} \frac{(O_{3j} - E_{3j})^2}{E_{3j}}$$

$$\sum_{j=1}^{c} \frac{(O_{1j} - E_{1j})^2}{E_{1j}} = \frac{(6-5{,}446809)^2}{5{,}446809} + \frac{(7-5{,}106383)^2}{5{,}106383} + \frac{(3-5{,}446809)^2}{5{,}446809} = 1{,}857552$$

$$\sum_{j=1}^{c} \frac{(O_{2j} - E_{2j})^2}{E_{2j}} = \frac{(5-5{,}106383)^2}{5{,}106383} + \frac{(3-4{,}787234)^2}{4{,}787234} + \frac{(7-5{,}106383)^2}{5{,}106383} = 1{,}371667$$

$$\sum_{j=1}^{c} \frac{(O_{3j} - E_{3j})^2}{E_{3j}} = \frac{(5-5{,}446809)^2}{5{,}446809} + \frac{(5-5{,}106383)^2}{5{,}106383} + \frac{(6-5{,}446809)^2}{5{,}446809} = 0{,}09505206$$

$$\chi^2 \approx 3{,}3243$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 < \chi^2_{0,95}(4)$ por lo que Fertility y Agriculture son independientes.

Fertility-Examination:

	Peor	Normal	Mejor	Sum
Baja	2	4	10	16
Media	7	3	5	15
Alta	11	5	0	16
Sum	20	12	15	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$E_{11} = E_{31} = \frac{20\times 16}{47} = 6{,}808511$$

$$E_{21} = \frac{20\times 15}{47} = 6{,}382979$$

$$E_{12} = E_{32} = \frac{12\times 16}{47} = 4{,}085106$$

$$E_{22} = \frac{12\times 15}{47} = 3{,}829787$$

$$E_{13} = E_{33} = \frac{15\times 16}{47} = 5{,}106383$$

$$E_{23} = \frac{15\times 15}{47} = 4{,}787234$$

$$\chi^2 = \sum_{j=1}^{c}\frac{(O_{1j}-E_{1j})^2}{E_{1j}} + \sum_{j=1}^{c}\frac{(O_{2j}-E_{2j})^2}{E_{2j}} + \sum_{j=1}^{c}\frac{(O_{3j}-E_{3j})^2}{E_{3j}}$$

$$\sum_{j=1}^{c}\frac{(O_{1j}-E_{1j})^2}{E_{1j}} = \frac{(2-6{,}808511)^2}{6{,}808511} + \frac{(4-4{,}085106)^2}{4{,}085106} + \frac{(10-5{,}106383)^2}{5{,}106383} = 8{,}0875$$

$$\sum_{j=1}^{c}\frac{(O_{2j}-E_{2j})^2}{E_{2j}} = \frac{(7-6{,}382979)^2}{6{,}382979} + \frac{(3-3{,}829787)^2}{3{,}829787} + \frac{(5-4{,}787234)^2}{4{,}787234} = 0{,}2488887$$

$$\sum_{j=1}^{c}\frac{(O_{3j}-E_{3j})^2}{E_{3j}} = \frac{(11-6{,}808511)^2}{6{,}808511} + \frac{(5-4{,}085106)^2}{4{,}085106} + \frac{(0-5{,}106383)^2}{5{,}106383} = 7{,}891666$$

$$\chi^2 \approx 16{,}228$$

Región Crítica: $\chi^2_{0{,}95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0{,}95}(4)$ por lo que Fertility y Examination son dependientes, entonces las calificaciones de los reclutas en los exámenes militares dependería de su fertilidad (directa o inversamente).

Fertility-Education:

	Mala	Normal	Buena	Sum
Baja	3	4	9	16
Media	8	4	3	15
Alta	10	2	4	16
Sum	21	10	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 8{,}2557$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 < \chi^2_{0,95}(4)$ por lo que Fertility y Education son independientes.

Fertility-Catholic:

	Protestantes	Mezclados	Católicos	Sum
Baja	6	9	1	16
Media	8	4	3	15
Alta	2	2	12	16
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 21{,}249$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0,95}(4)$ por lo que Fertility y Catholic son dependientes. Podemos ver como las provincias católicas tienen mayor fertilidad que las protestantes y mezcladas.

Fertility-Infant.Mortality:

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.

	Baja	Media	Alta	Sum
Baja	9	3	4	16
Media	4	7	4	15
Alta	3	5	8	16
Sum	16	15	16	47

- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 7{,}3715$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 < \chi^2_{0,95}(4)$ por lo que Fertility e Infant.Mortality son independientes.

Agriculture-Examination:

	Peor	Normal	Mejor	Sum
Urbanitos	3	3	10	16
Agr_Urb	5	6	4	15
Agricultores	12	3	1	16
Sum	20	12	15	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 16{,}316$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0,95}(4)$ por lo que Agriculture y Examination son dependientes. Podemos ver como las provincias de agricultores tienen peor calificaciones que las urbanas las cuales tienen la mejor calificación.

Agriculture-Education:

	Mala	Normal	Buena	Sum
Urbanitos	5	2	9	16
Agr_Urb	3	5	7	15
Agricultores	13	3	0	16
Sum	21	10	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 17{,}599$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0,95}(4)$ por lo que Agriculture y Education son dependientes. Podemos ver que ninguna provincia agricultora tiene buena educación, y la mayoría de estas tiene una mala educación.

Agriculture-Catholic:

	Protestantes	Mezclados	Católicos	Sum
Urbanitos	4	10	2	16
Agr_Urb	7	4	4	15
Agricultores	5	1	10	16
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 15{,}472$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0,95}(4)$ por lo que Agriculture y Catholic son dependientes. Podemos ver que provincias de agricultores tienen una mayor población católica mientras que las provincias urbanas tienen más diversidad en creencias.

Agriculture-Infant.Mortality:

	Baja	Media	Alta	Sum
Urbanitos	3	9	4	16
Agr_Urb	7	1	7	15
Agricultores	6	5	5	16
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 8{,}948$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 < \chi^2_{0,95}(4)$ por lo que Agriculture e Infant.Mortality son independientes.

Examination-Education:

	Mala	Normal	Buena	Sum
Peor	15	4	1	20
Normal	3	4	5	12
Mejor	3	2	10	15
Sum	21	10	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 18{,}33$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0,95}(4)$ por lo que Examination y Education son dependientes. Podemos ver claramente que para provincias con peor nivel de eduación las calificaciones son peores y para las provincias con mejor nivel de educación las calificaciones son mejores.

Examination-Catholic:

	Protestantes	Mezclados	Católicos	Sum
Peor	5	1	14	20
Normal	6	4	2	12
Mejor	5	10	0	15
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 25{,}37$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0,95}(4)$ por lo que Examination y Catholic son dependientes. Podemos ver que para provincias católicas están la mayoría de peores calificaciones y ninguna tiene buenas calificaiones, luego las provincias con más diversidad en creencias son las que más tienen mejores calificaciones.

Examination-Infant.Mortality:

	Baja	Media	Alta	Sum
Peor	6	6	8	20
Normal	2	4	6	12
Mejor	8	5	2	15
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 5{,}8358$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 < \chi^2_{0,95}(4)$ por lo que Examination e Infant.Mortality son independientes.

Education-Catholic:

	Protestantes	Mezclados	Católicos	Sum
Mala	7	3	11	21
Normal	5	2	3	10
Buena	4	10	2	16
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 12{,}619$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 > \chi^2_{0,95}(4)$ por lo que Education y Catholic son dependientes. Como la mayoría de provincias católicas son de agricultores y estas tienen una mala educación, se puede apreciar que en el caso de la mayoría de provinicas católicas existe una peor educación, mientras que las provincias con más diversidad en creencias tienen mayor mejor educación.

Education-Infant.Mortality:

	Baja	Media	Alta	Sum
Mala	6	8	7	21
Normal	3	3	4	10
Buena	7	4	5	16
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 1{,}3221$$

Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 < \chi^2_{0,95}(4)$ por lo que Education e Infant.Mortality son independientes.

Catholic-Infant.Mortality:

	Baja	Media	Alta	Sum
Protestantes	5	6	5	16
Mezclados	6	5	4	15
Católicos	5	4	7	16
Sum	16	15	16	47

Prueba Chi-Cuadrado de Independencia:

Verifiquemos si existe relación entre ambas variables categóricas.
Hipótesis:

- **Hipótesis Nula** (H_0): Hay independencia entre las variables.
- **Hipótesis Alternativa** (H_1): Hay dependencia entre las variables.

Estadístico de Prueba:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 1{,}3545$$

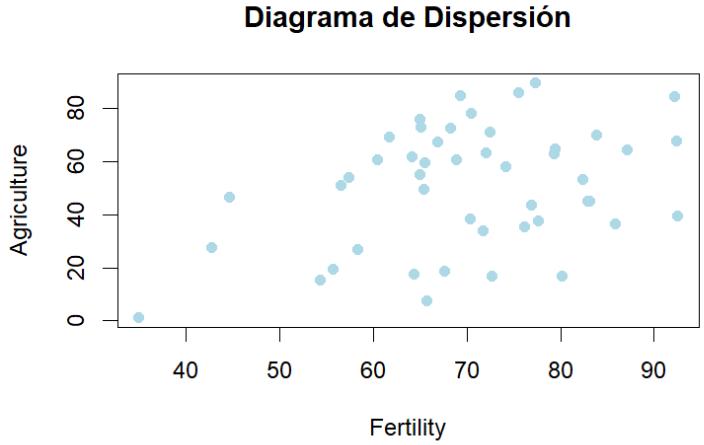
Región Crítica: $\chi^2_{0,95}(4) = 9{,}487729$

Luego, $\chi^2 < \chi^2_{0,95}(4)$ por lo que Catholic e Infant.Mortality son independientes.

Análisis de correlación entre las variables:

Fertility-Agriculture:

Hallamos el coeficiente de correlación de Spearman, dado que Fertility sigue una distribución normal, sin embargo Agriculture no se puede afirmar con certeza de que siga una distribución normal, por lo que a pesar de ser ambas valores cuantitativos, no podemos usar Pearson si al menos una de las variables no distribuye normal.
r = 0.2426643
Este valor pertenece al intervalo [-0.4, 0.4] que define la no existencia de correlación lineal entre el procentaje de fertilidad y agricultores en las provincias.



Fertility-Examination:

En este caso, ambas variables distribuyen normal, pero al Fertility tener valores atípicos es mejor usar Spearman.
r = -0.66090300

Este valor pertenece al intervalo [-0.7, -0.4] el cual no define si existe correlación lineal entre el porcentaje de fertilidad y buenas calificaciones entre los reclutas. Sin embargo podemos apreciar que está más cercano al extremo del intervalo que define la existencia de correlación. Dado el

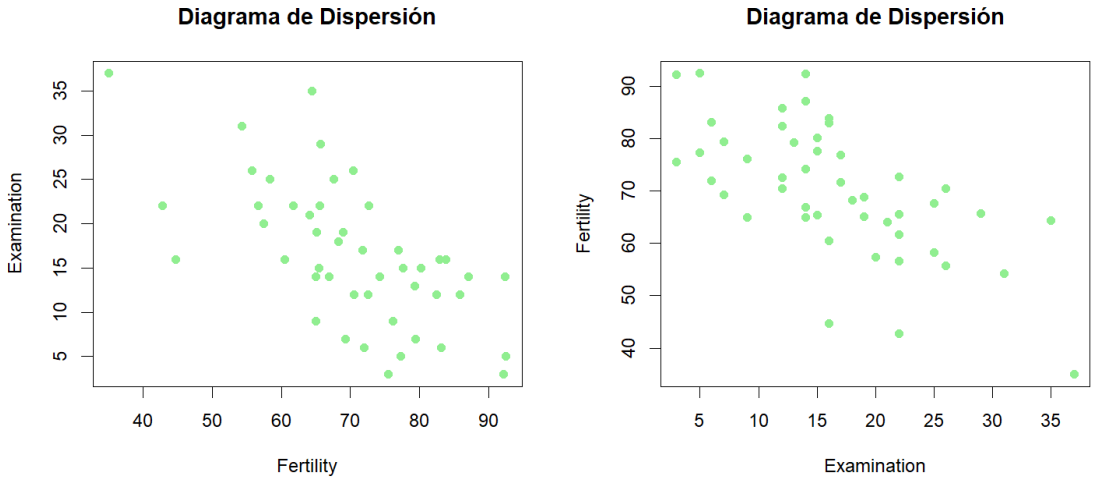
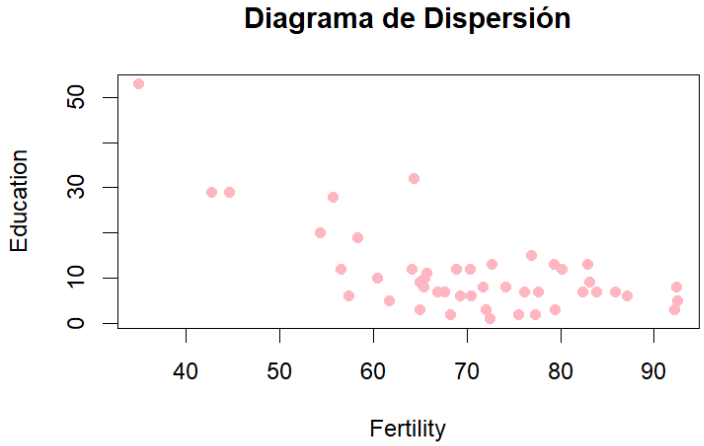


diagrama de dispersión de ambos, podemos notar un poco de dispersión de los datos, sin embargo siguen cierto patrón de dispersión el cual no define una correlación lineal entre ambas variables, pero sí una posible correlación cuadrática.

Fertility-Education:

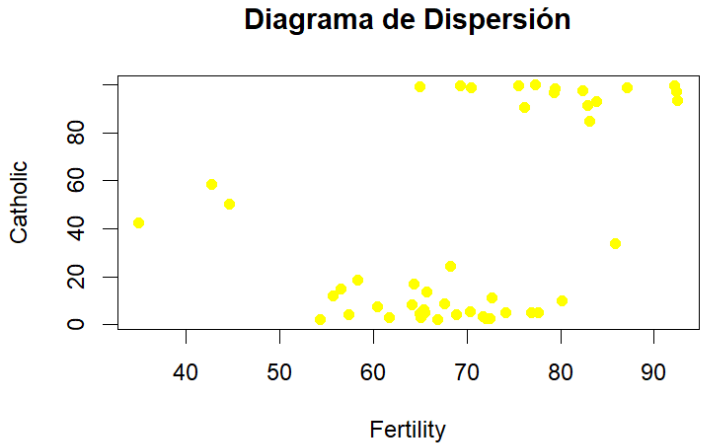
En este caso, Education no distribuye normal, por lo que es necesario usar el coeficiente de correlación de Spearman.
r = -0.4432577
Este valor pertenece al intervalo [-0.7, -0.4] el cual no define si existe correlación lineal entre el porcentaje de fertilidad y buena educación entre los reclutas. Sin embargo podemos apreciar que está más cercano al extremo del intervalo que define la no existencia de correlación lineal. Dado

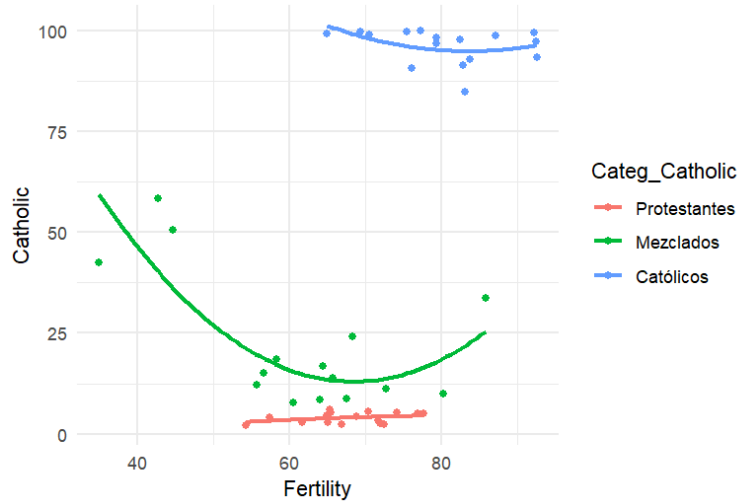


el diagrama de dispersión de ambos, podemos notar un poco de dispersión de los datos, sin embargo siguen cierto patrón de dispersión el cual puede definir una correlación cuedrática entre ambas variables.

Fertility-Catholic:

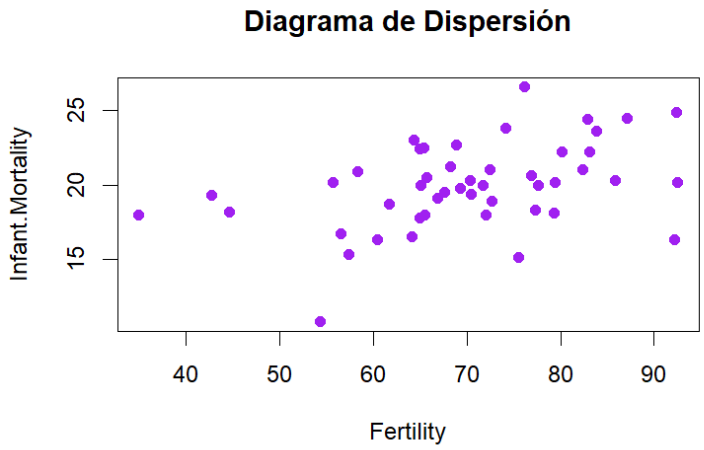
En este caso, Catholic no distribuye normal, por lo que es necesario usar el coeficiente de correlación de Spearman.
$r = 0.4136456$
Este valor pertenece al intervalo [0.4, 0.7] el cual no define si existe correlación lineal entre el porcentaje de fertilidad y el catolicismo. Sin embargo podemos apreciar que está más cercano al extremo del intervalo que define la no existencia de correlación lineal, y esta es corroborada por el diagarama de dispersión de ambas variables.





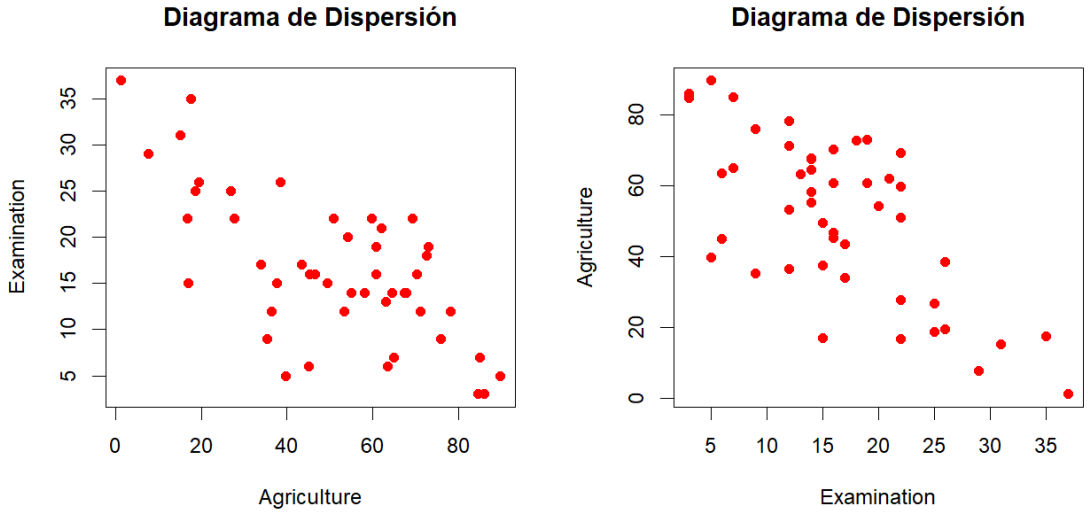
Fertility-Infant.Mortality:

En este caso, ambas distribuyen normal, por lo que es seguro utilizar el coeficiente de correlación de Pearson.
$r = 0.416556$
Este valor pertenece al intervalo [0.4, 0.7] el cual no define si existe correlación lineal entre el porcentaje de fertilidad y la mortalidad infantil. Sin embargo podemos apreciar que está más cercano al extremo del intervalo que define la no existencia de correlación lineal, y esta es corroborada por el diagarama de dispersión de ambas variables.



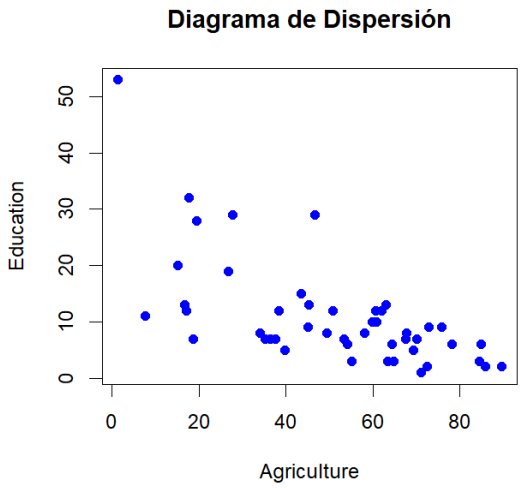
Agriculture-Examination:

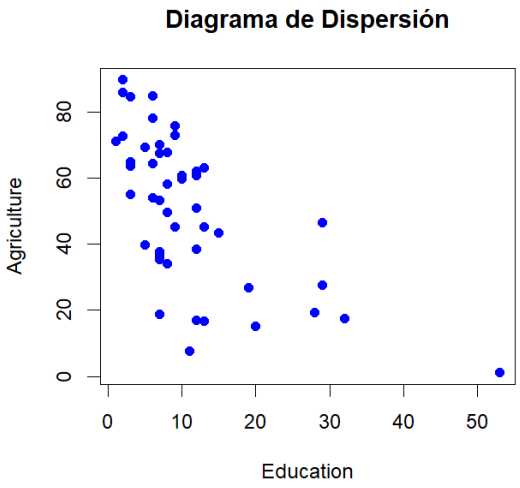
En este caso, como no podemos afirmar que Agriculture siga una distribución normal, aplicaremos para todos sus análsis de correlación el coeficiente de correlación de Spearman. r = -0.5988599 Este valor pertenece al intervalo [-0.7, -0.4] el cual no define si existe correlación lineal entre el porcentaje de agricultores y las buenas calificaciones de los reclutas, sin embargo el valor es más cercano al extremo del intervalo que define una correlación lineal inversa. Esto se puede corroborar con el patrón de dispersión que siguen los valores de nuestras variables.



Agriculture-Education:

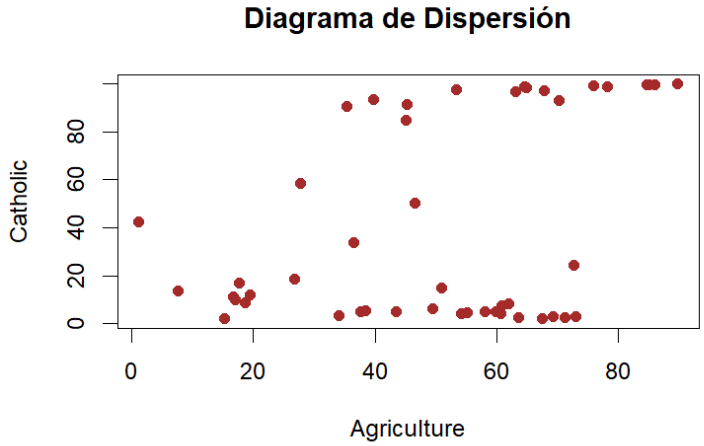
r = -0.6504638
Este valor pertenece al intervalo [-0.7, -0.4] el cual no define si existe correlación lineal entre el porcentaje de agricultores y la educuación de los reclutas, sin embargo el valor es más cercano al extremo del intervalo que define una correlación lineal inversa. Esto se puede corroborar con el patrón de dispersión que siguen los valores de nuestras variables. También podría tratarse de una correlación logarítmica o cuadrática dado el patrín que siguen las variables.

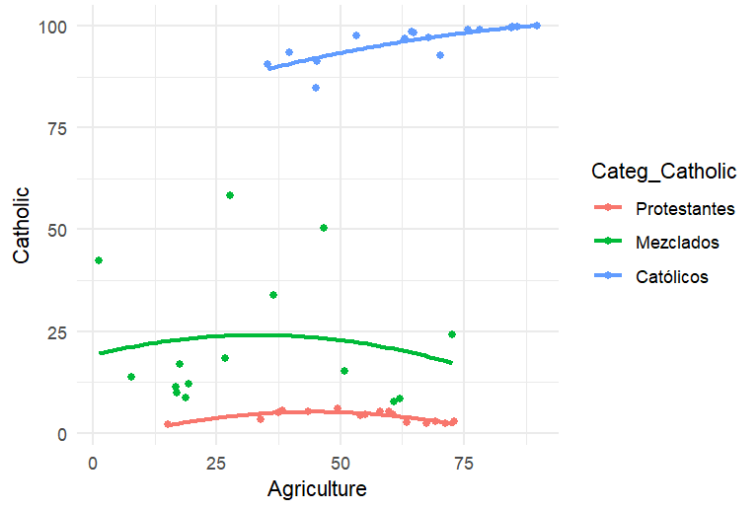




Agriculture-Catholic:

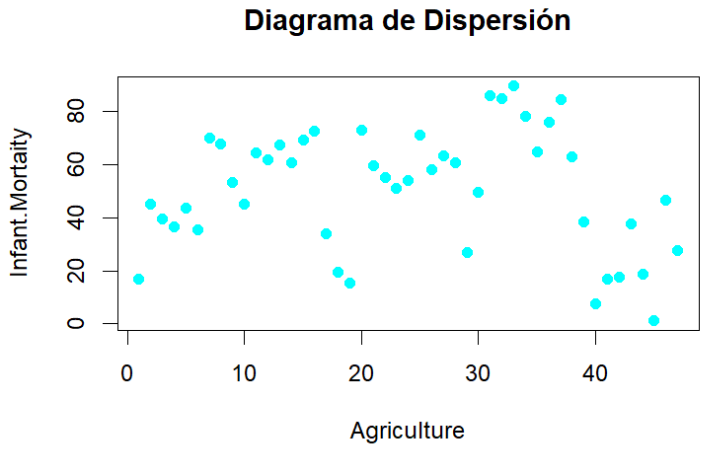
$r = 0.2886878$
Este valor pertenece al intervalo [0, 0.4] el cual define la no existencia de correlación lineal entre el porcentaje de agricultores y católicos.





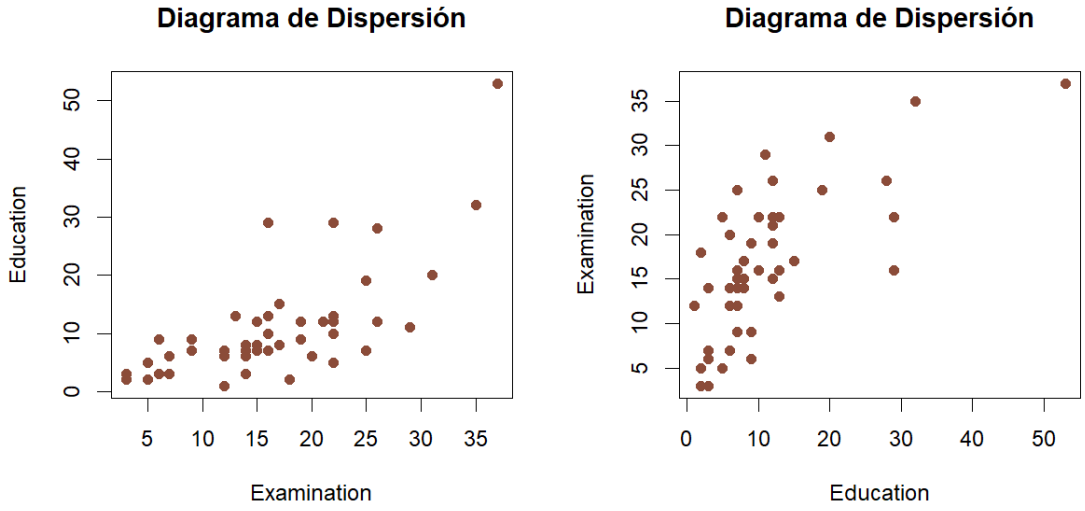
Agriculture-Infant.Mortality:

r = -0.1521287
Este valor pertenece al intervalo [-0.4, 0] el cual define la no existencia de correlación entre el porcentaje de agricultores y la mortalidad infantil.



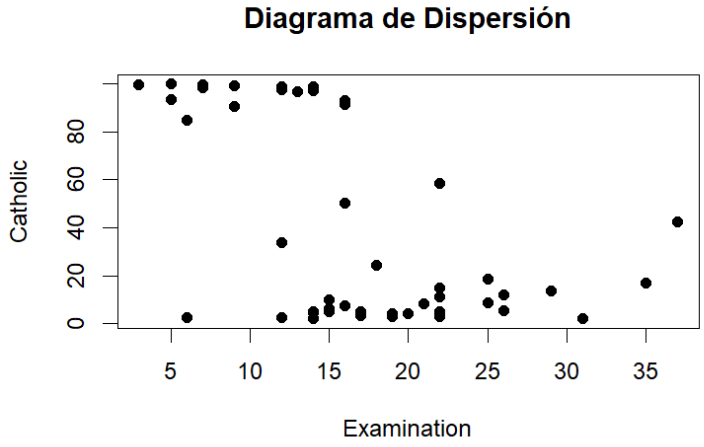
Examination-Education:

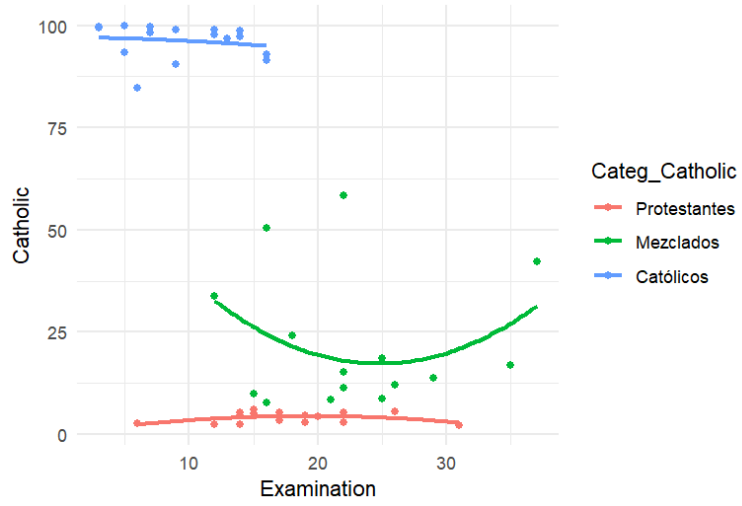
En este caso, Examination sigue una distribución normal, pero Education no, por lo que se utiliza el coeficiente de correlación de Spearman.
r = 0.6746038
Este valor pertenece al intervalo [0.4, 0.7] el cual no define si existe correlación lineal entre el porcentaje de buenas calificaciones y buena educación en los reclutas. Luego este valor es más cercano al extremo de existencia de correlación lineal directa, cosa que se puede apreciar el patrón que siguen los valores en el diagrama de dispersión. Podría también tratarse de una relación logarítmica en un diagrama y cuadrática en el otro.



Examination-Catholic:

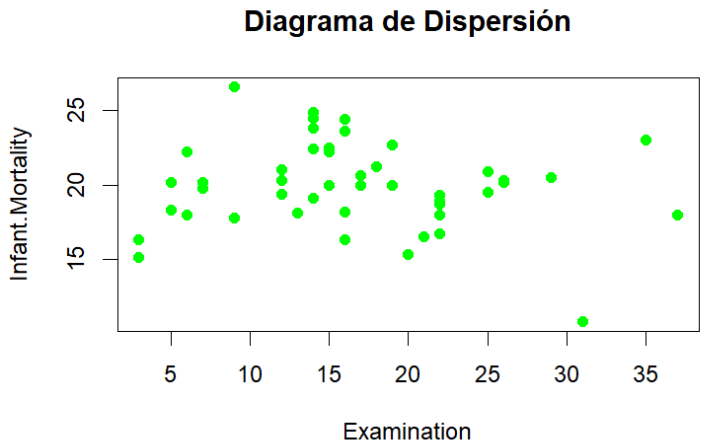
En este caso, Examination sigue una distribución normal, pero Catholic no, por lo que se utiliza el coeficiente de correlación de Spearman.
r = -0.4750575
Este valor pertenece al intervalo [-0.7, -0.4] el cual no define si existe correlación entre el porcentaje de reclutas con altas notas y el catolicismo, luego este valor es más ceracno al extremo del intervalo que define la no correlación. Esto se puede corroborar con el diagrama de dispersión.





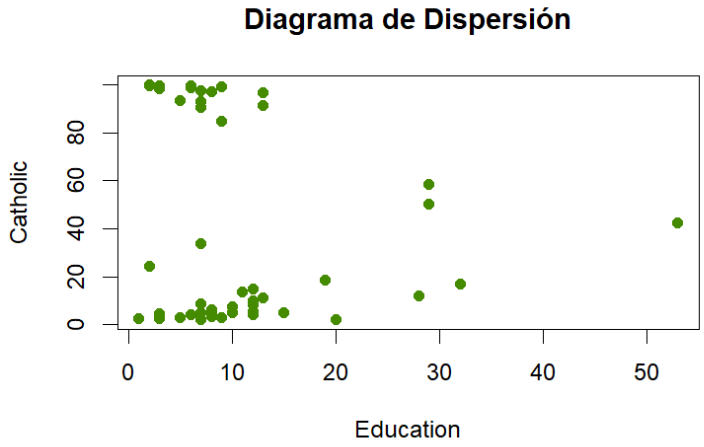
Examination-Infant.Mortality:

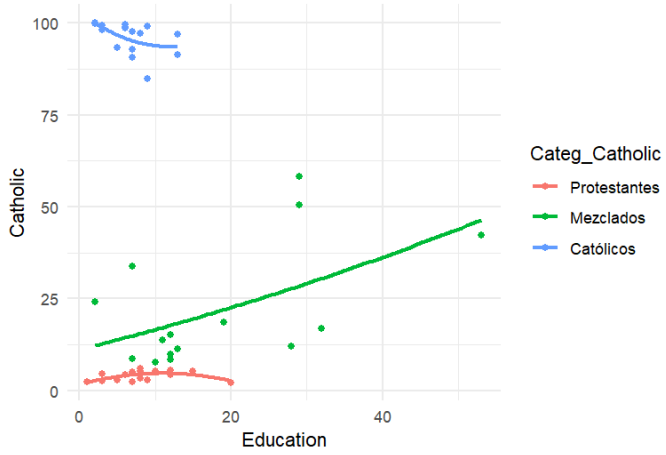
En este caso ambas distribuyen normal por lo que es seguro usar coeficiente de correlación de Pearson.
r = -0.1140216
Este valor pertenece al intervalo [-0.4, 0] el cual define la no existencia de correlación entre el procentaje de reclutas con altas calificaciones y la mortalidad infantil.



Education-Catholic:

En este caso Education no disribuye normal, por lo que se aplicara coeficiente de correlación de Spearman para todos sus análisis de correlación.
r = -0.1444163
Este valor pertenece al intervalo [-0.4, 0] el cual define la no existencia de correlación entre el procentaje de reclutas con mejor educación y el catolicismo. Podemos ver que no importa la posición

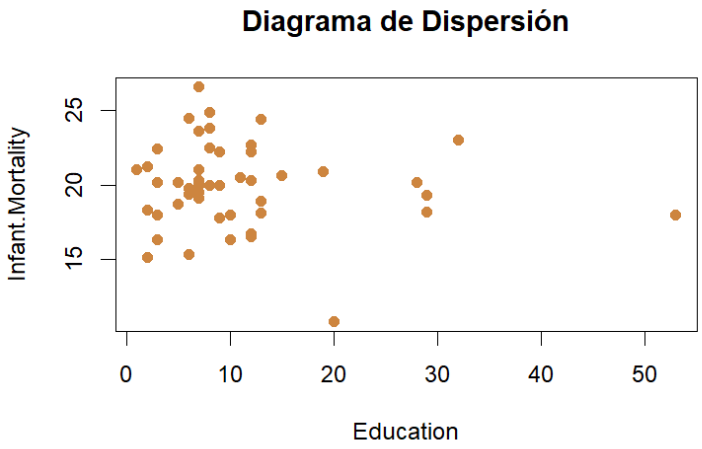




religiosa, el nivel educacional por lo general es bastante bajo y no están relacionados.

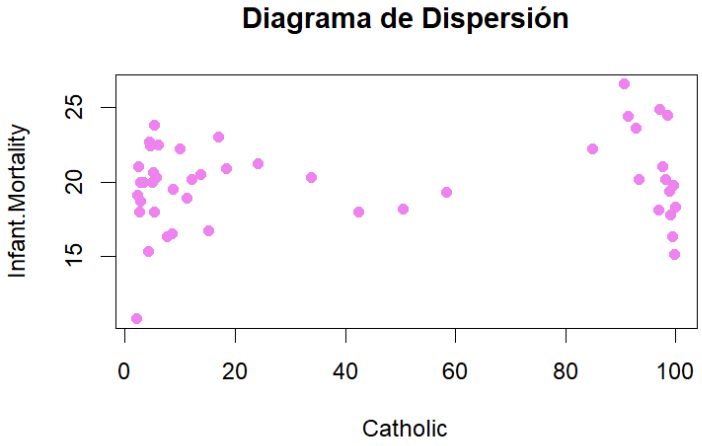
Education-Infant.Mortality:

r = -0.01898137
Este valor pertenece al intervalo [-0.4, 0] el cual define la no existencia de correlación entre el procentaje de reclutas con mejor educación y la mortalidad infantil.



Catholic-Infant.Mortality:

En este caso Catholic no disribuye normal, por lo que se aplicara coeficiente de correlación de Spearman para todos sus análisis de correlación.
r = 0.06611714
Este valor pertenece al intervalo [0, 0.4] el cual define la no existencia de correlación entre el procentaje de católicos y la mortalidad infantil.



Regresión Lineal:

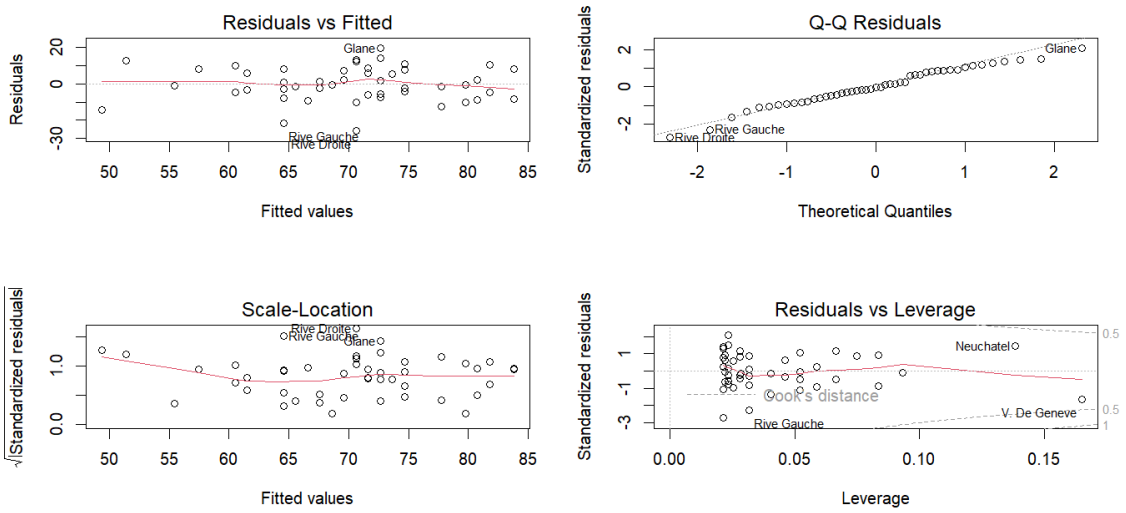
Fertility:

La única variable con una posible relación lineal inversa con la fertilidad de las provinicas es el porcentaje de altas calificaciones en los recultas militares. En qué afectará Examination a Fertility?.

Planteando el modelo de regresión lineal simple:
$Fertility = 86{,}8185 + Examination * (-1{,}0113)$

```
Call:
lm(formula = Fertility ~ Examination, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9375  -6.0044  -0.3393   7.9239  19.7399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.8185     3.2576  26.651  < 2e-16 ***
Examination  -1.0113     0.1782  -5.675 9.45e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.642 on 45 degrees of freedom
Multiple R-squared:  0.4172,    Adjusted R-squared:  0.4042
F-statistic: 32.21 on 1 and 45 DF,  p-value: 9.45e-07
```

Podemos observar que el modelo no es bueno, dado que el rango de residuo es bastante grande al igual que el error estándar de los residuos. Sin embargo la variable Examination tiene una gran significancia a la hora de predecir la fertilidad en las provincias, sin embargo el modelo solo explica el 0.4172 de la variación en la fertilidad, por lo que de nuevo, el modelo es muy malo y no tenemos suficiente información para hacer una predicción correcta.



Podemos ver que se cumple la linealidad, la normalidad de los residuos con un p-value igual a 0.6071 en el test de Shapiro-Wilk, la media(3.918999e-16) y suma(1.84297e-14) de los residuos tiende a 0, la homocedasticidad con un p-value igual a 0.6192 en el test de Breusch-Pagan y finalmente no cumplen la independencia de los errores ya que el test de Durbin-Watson nos da un p-value igual a 1.759e-08 <0.05, por lo que este modelo no cumple con todos los supuestos del modelo.

Examination:

Las variables que parecen terner cierta relación lineal con Examination son:

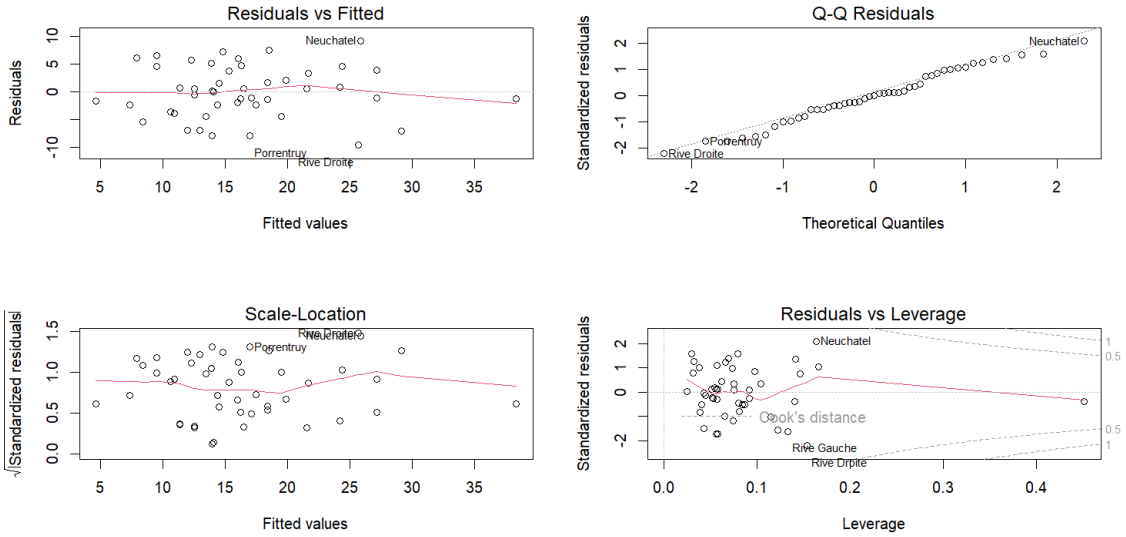
- Fertility (lineal inversa)
- Agriculture (lineal inversa)

- Education (lineal directa)

Planteando el modelo de regresión lineal múltiple:
$Examination = 40{,}31441 + Fertility * (-0{,}24500) + Agriculture * (-0{,}15887) + Education * (0{,}12823)$

```
Call:
lm(formula = Examination ~ Agriculture + Fertility + Education,
    data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6784 -2.4231  0.0687  3.7547  9.1561

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.31441    7.00702   5.753 8.34e-07 ***
Agriculture -0.15887    0.04057  -3.916 0.000317 ***
Fertility   -0.24500    0.07582  -3.231 0.002367 **
Education    0.12823    0.11988   1.070 0.290745
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.767 on 43 degrees of freedom
Multiple R-squared:  0.6662,    Adjusted R-squared:  0.6429
F-statistic: 28.61 on 3 and 43 DF,  p-value: 2.498e-10
```

Podemos ver que el modelo parece ser relativamengte bueno, a pesar de que solo explica el 0.6429 de la variación de los porcentajes de reclutas con buenas calificaciones en el servicio militar. El rango de residuo es relativamente pequeño, con una media de los residuos crecana a cero. Education podemos ver que no está aportando nada al modelo, a pesar de la relación directa que tiene el nivel educacional con las buenas calificaciones.

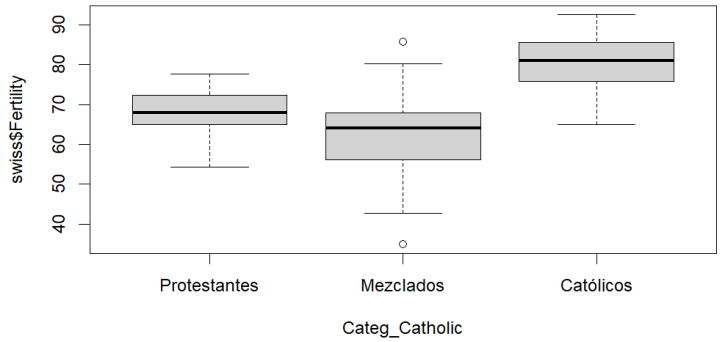


Podemos ver que se cumple la linealidad, la nomralidad de los residuos con un p-value igual a 0.5802 en el test de Shapiro-Wilk, la media(-2.220446e-16) y suma(-1.04361e-14) de los errores tienden a cero, la homocedasticidad con un p-value igual a 0.6044 en el test de Breusch-Pagan y finalmente tampoco cumple la independencia de los errores ya que el test de Durbin-Watson nos da un p-value igual a 0.0004043 <0.05, por lo que este modelo no cumple con todos los supuestos del modelo.

Análisis de Varianza:

Existen diferencias significativas en cuanto a la fertilidad de los diferentes grupos religiosos de las provincias?:

Podemos ver en el gráfico que es posible que exista un efecto del grupo religioso en la fertilidad de las provincias.



Realizando ANOVA obtenemos los siguientes resultados:

```
              Df Sum Sq Mean Sq F value   Pr(>F)
Categ_Catholic  2   2940  1469.9   15.26 9.25e-06 ***
Residuals      44   4238    96.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

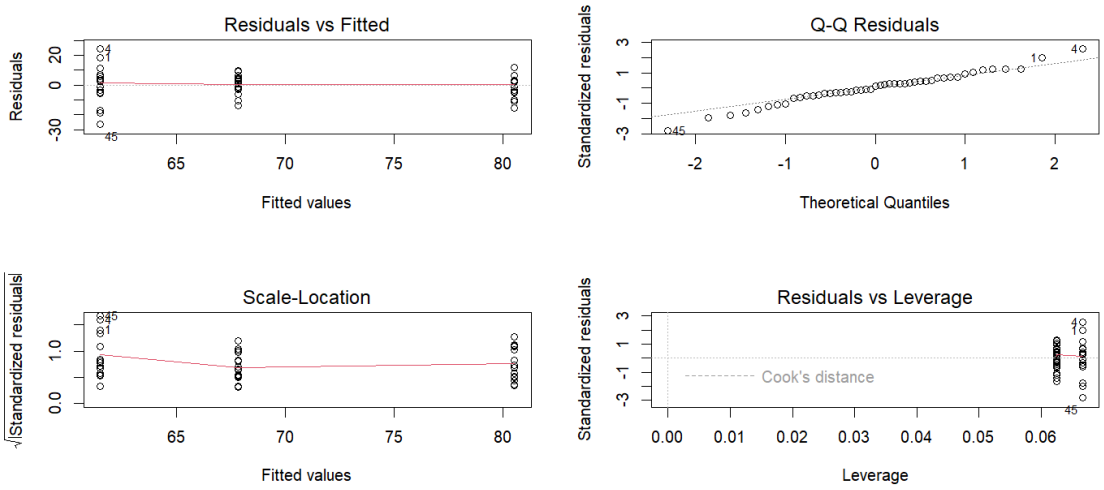
A pesar de que lógicamente mayor parte de la variabilidad de la fertilidad de las provincias no se explica por los grupos religiosos, sí que podemos ver que el p-value igual a 9.25e-06 <0.05, se acepta de que hay diferencias estadísticamente significativas en la fertilidad entre al menos dos de los grupos de religión.
Evaluando las diferencias específicas:

```
$Categ_Catholic
                              diff        lwr      upr     p adj
Mezclados-Protestantes  -6.350417 -14.905793  2.20496 0.1812898
Católicos-Protestantes  12.706250   4.289994 21.12251 0.0018975
Católicos-Mezclados     19.056667  10.501290 27.61204 0.0000075
```

- Mezclados vs. Protestantes: No hay evidencia suficiente de que Mezclados y Protestantes tengan fertilidades diferentes.
    - Diferencia media: -6.35 (Mezclados tienen menor Fertility que Protestantes).
    - p-value = 0.1813 (no significativo, p >0.05).
- Católicos vs. Protestantes: Los Católicos tienen una fertilidad significativamente mayor que los Protestantes.
    - Diferencia media: 12.71 (Católicos tienen mayor Fertility que Protestantes).
    - p-value = 0.0019 (significativo, p <0.05).

- Católicos vs. Mezclados: Los Católicos tienen una fertilidad significativamente mayor que los Mezclados.
    - Diferencia media: 19.06 (Católicos tienen mayor Fertility que Mezclados).
    - p-value = 0.0000075 (altamente significativo, p <0.001).

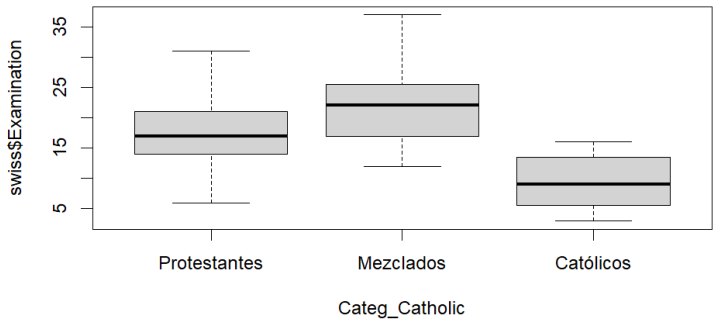
Verificando que se cumplan los supuests del modelo:



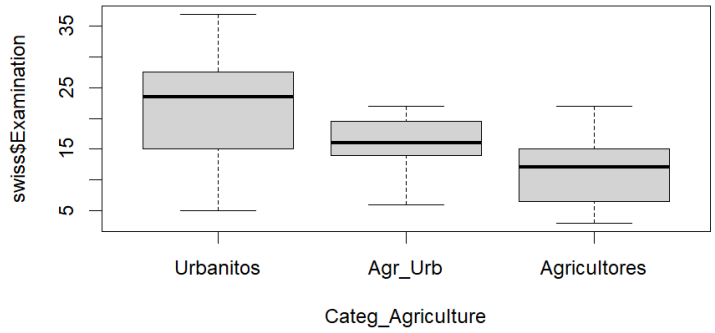
Se cumple la nomralidad de los residuos, con un p-value igual a 0.6133 >0.05 en el test de Shapiro-Wilk, no se cumple la homogeneidad de las varianzas con un p-value igual a 0.01656 <0.05 en el test de Bartlett, y por ultimo no se cumple la independencia de las varianzas, con un p-value igual a 0.0003585 <0.05 en el test de Durbin-Watson. Por lo que este modelo no es válido.

Existen diferencias significativas en cuanto al porcentaje de reclutas militares con altas notas en el servicio militar en provincias con determinado nivel educativo y grupo religioso?:

Podemos ver en el gráfico que es posible que exista un efecto del grupo religioso en el porcentaje de reclutas militares con altas notas en el servicio militar.



Podemos ver en el gráfico que es posible que exista un efecto del nivel educativo de las provincias en el porcentaje de reclutas militares con altas notas en el servicio militar.



Realizando ANOVA obtenemos los siguientes resultados:

```
                Df Sum Sq Mean Sq F value   Pr(>F)
Categ_Catholic   2 1403.4   701.7  24.067 1.09e-07 ***
Categ_Education  2  299.7   149.9   5.139   0.0101 *
Residuals       42 1224.6    29.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El efecto de Categ_Catholic en Examination es muy significativo, con un p-value igual a 1.09e-07 ($p < 0.001$). La religión tiene un efecto fuerte en los puntajes de Examination. El efecto de Categ_Education en Examination es significativo, con un p-value igual a 0.0101 ($p < 0.05$). La categoría de educación también influye en los puntajes de Examination, pero con menor impacto que la religión.
Evaluando las diferencias específicas:

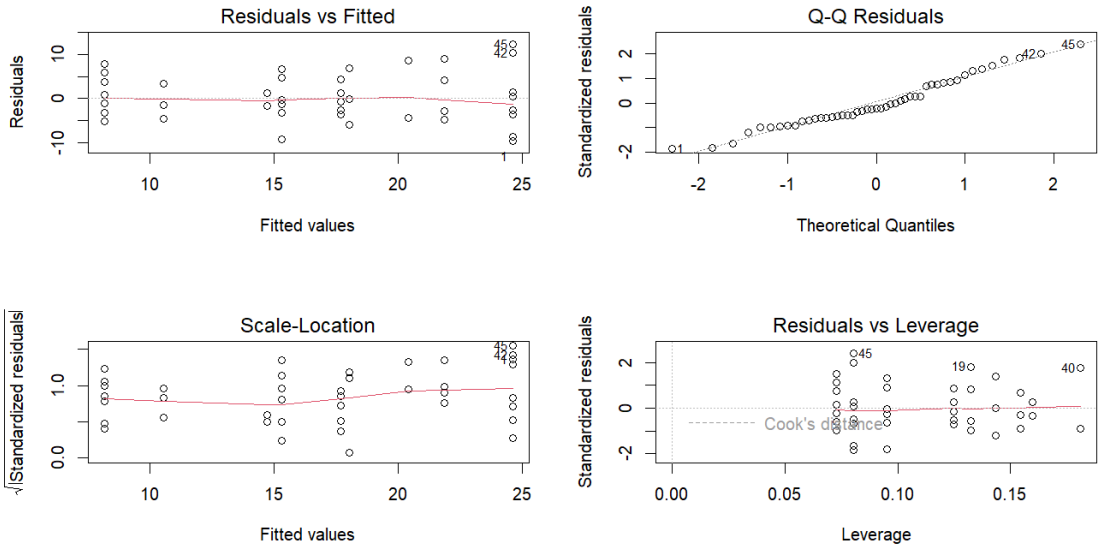
```
$Categ_Catholic
                            diff         lwr       upr     p adj
Mezclados-Protestantes   5.045833   0.3310178  9.760649 0.0335783
Católicos-Protestantes  -8.250000 -12.8881468 -3.611853 0.0002697
Católicos-Mezclados    -13.295833 -18.0106489 -8.581018 0.0000001

$Categ_Education
                 diff        lwr      upr     p adj
Normal-Mala  2.012857 -3.0274927 7.053207 0.5995060
Buena-Mala   5.012128  0.6588068 9.365449 0.0207266
Buena-Normal 2.999271 -2.2890302 8.287572 0.3614613
```

- Conclusión sobre Categ_Catholic:
  - Los Mezclados tienen un puntaje significativamente mayor que los Protestantes ($p = 0.0336 < 0.05$).
  - Los Católicos tienen un puntaje significativamente menor que los Protestantes ($p = 0.00027 < 0.001$).
  - Los Católicos tienen un puntaje aún más bajo que los Mezclados ($p = 0.0000001 < 0.001$).
  - El grupo con mayor puntaje en Examination es Mezclados, y el más bajo es Católicos.
- Conclusión sobre Categ_Education:
  - La provincias con buena educación tiene un puntaje significativamente mayor que las de mala educación ($p = 0.0207 < 0.05$).
  - No hay diferencias significativas entre las provincias con nivel educativo normal y malo ni entre bueno y normal.

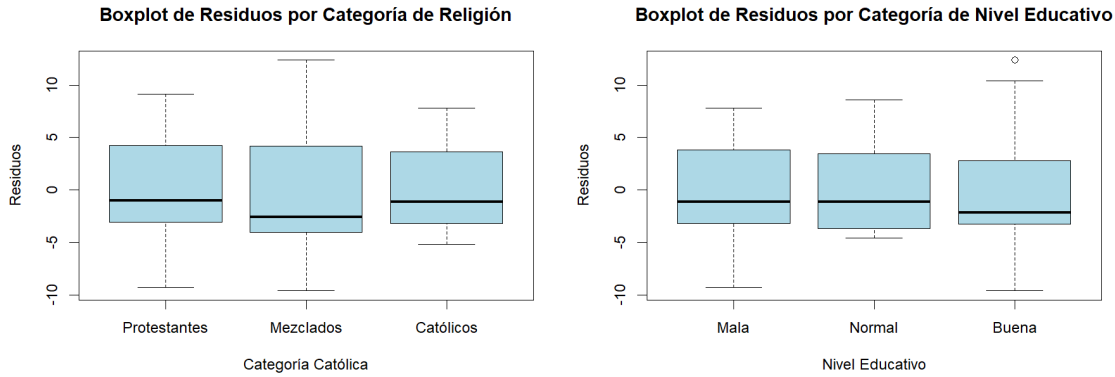
- Tener una buena educación mejora significativamente el puntaje en Examination, pero el paso de mala a normal no es suficiente para que la diferencia sea significativa.

- El efecto de la religión parece ser más fuerte que el de la educación en los resultados del examen.

Verificando que se cumplan los supuests del modelo:



Se incumple la normalidad de lor residuos con un pvalue igual a 0.2038 <0.05 en el test de Shapiro-Wilk, se cumple la independencia de las varianzas con un p-value igual a 0.05424 >0.05 en el test de Durbin-Watson, y por ultimo podemos ver dada la figura de arriba y la siguiente que no se cumple la homogeneidad de las varianzas, por lo que este modelo tampoco es válido.



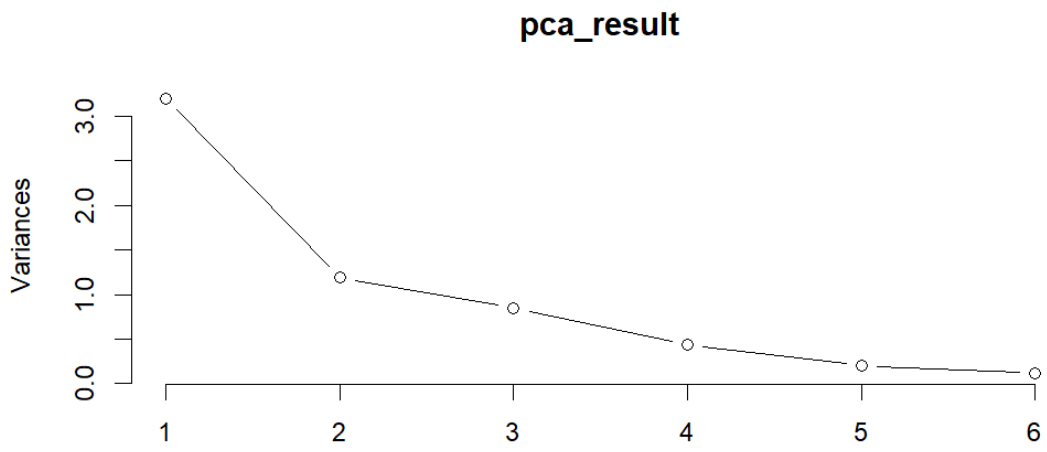
Análisis de componentes principales:

Veamos si con este análisis podemos sacar alguna otra relación entre las variables con este análisis dado que al tener solo 6 de ellas no es necesario reducirlo.
Realizando el análisis:

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.7888 1.0901 0.9207 0.66252 0.45225 0.34765
Proportion of Variance 0.5333 0.1981 0.1413 0.07315 0.03409 0.02014
Cumulative Proportion  0.5333 0.7313 0.8726 0.94577 0.97986 1.00000
```

- PC1 tiene la mayor desviación estándar (1.7888), lo que significa que capta la mayor variabilidad de los datos.
- PC1 explica el 53.33 % de la variabilidad total.
- PC2 añade un 19.81 % más, es decir, con PC1 + PC2 ya explicamos el 73.13 % de la variabilidad total.
- PC3 aporta un 14.13 %, y con los tres primeros ya explicamos el 87.26 % de la variabilidad total.

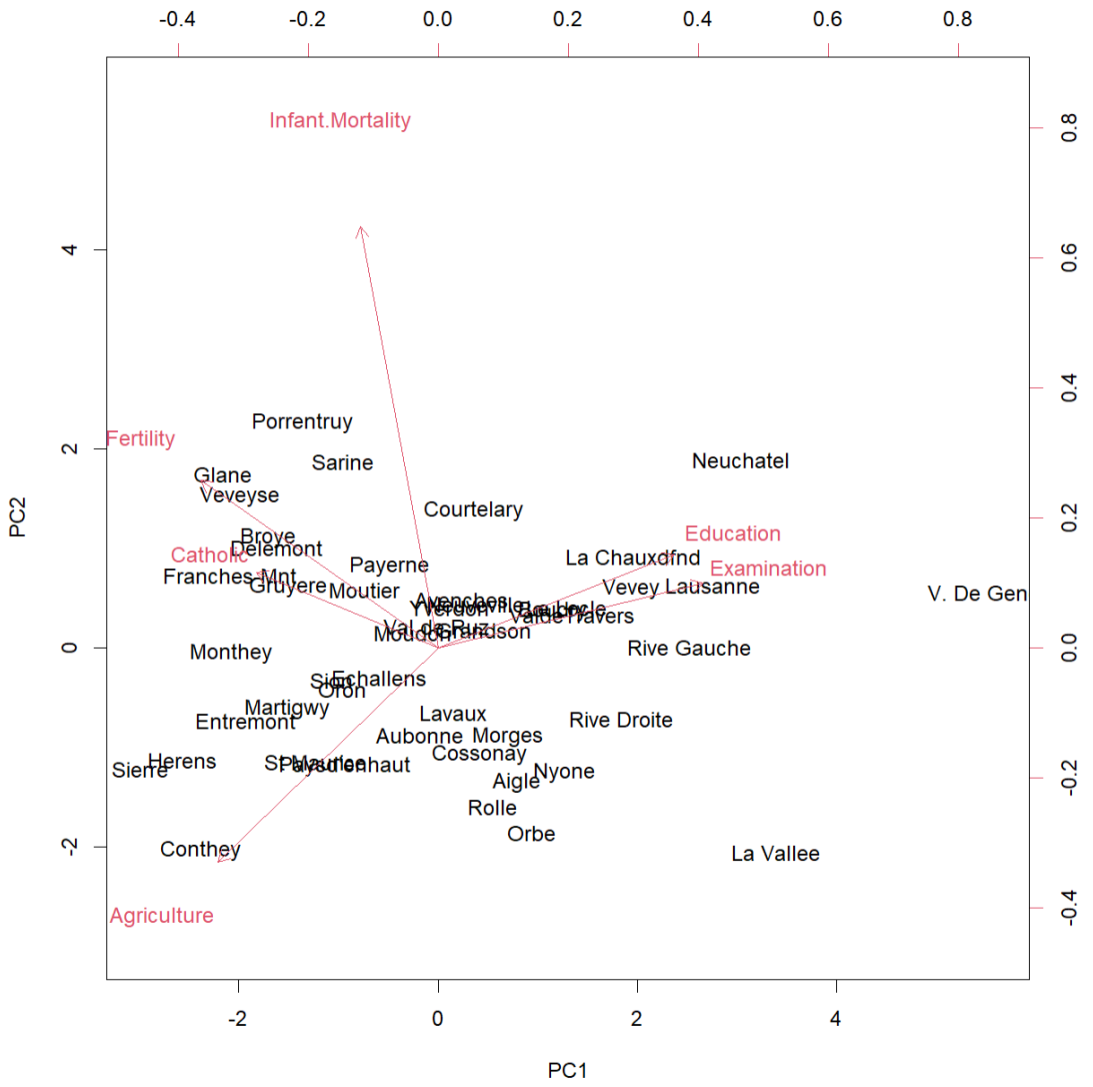
Como se puede ver en la siguiente gráfica los últimos componentes aportan cada vez menos información:



Veamos la matriz de rotación que nos dice cómo están relacionadas nuestras variables originales con nuestros componentes principales:

	PC1	PC2	PC3	PC4	PC5	PC6
Fertility	-0.4569876	0.3220284	-0.17376638	0.53555794	0.38308893	-0.47295441
Agriculture	-0.4242141	-0.4115132	0.03834472	-0.64291822	0.37495215	-0.30870058
Examination	0.5097327	0.1250167	-0.09123696	-0.05446158	0.81429082	0.22401686
Education	0.4543119	0.1790495	0.53239316	-0.09738818	-0.07144564	-0.68081610
Catholic	-0.3501111	0.1458730	0.80680494	0.09947244	0.18317236	0.40219666
Infant.Mortality	-0.1496668	0.8111645	-0.16010636	-0.52677184	-0.10453530	0.07457754

PC1 captura que zonas con alta educación y exámenes tienden a tener menos fertilidad y menos dependencia de la agricultura dado que existe un contraste entre educación/modernidad vs. fertilidad/agricultura. PC2 parece estar relacionado con la mortalidad infantil y la fertilidad, indicando que en ciertas regiones la mortalidad infantil es alta junto con la fertilidad. Y la PC3 parece captar que lugares con un alto porcentaje de católicos parecen tener un patrón particular en relación con la educación ya que esta contibuye, pero en una dirección diferente que Catholic.



Podemos ver en el gráfico que las provincias V. De Geneve, La Vallee y Neuchatel tienen altos niveles de educación y reclutas con buenas calificaciones en los exámenes militares, baja fertilidad y poca dependencia de la agricultura. Estas probablemente son áreas urbanas con economías modernas y diversificadas.
La provincias Sierre, Herens y Conthey tienen mayor dependencia de la agricultura, tasas de fertilidad más altas y niveles de educación y examinación más bajos. Probablemente son áreas más rurales y tradicionales