Taller 1:Diseño de experimentos

Sergio Andres Diaz Vera

2022-08-23

Análisis y Comparación

En un estudio realizado en India en mayo de 2021 se quiería saber si hay diferencia significativa en la carga viral entre personas infectadas con Sars-Cov-2 con la variante beta y la variante delta.

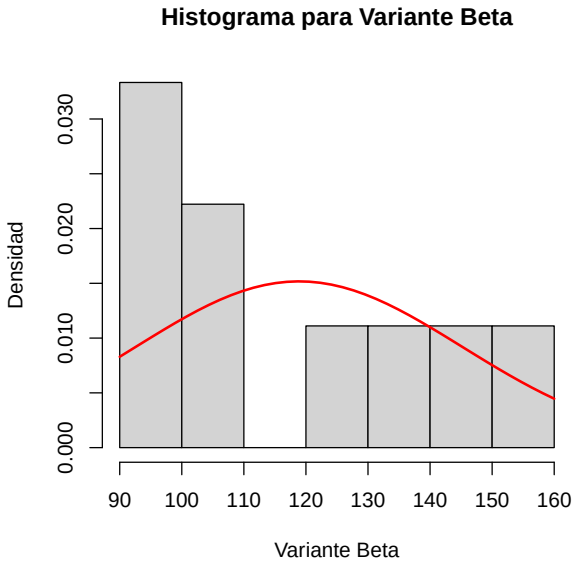
1. Se procede a observar la carga viral entre personas infectadas con la variante beta y delta

	1	2	3	4	5	6	7	8	9
Variante Beta	100	102	130	140	150	160	90	103	95

	1	2	3	4	5	6	7	8	9	10	11
Variante Delta	128	156	100	98	120	160	150	120	129	137	140

Análisis descriptivo Variante Beta

Es posible observar , aunque con pocos datos , la variante beta posee un comportamiento normal además de no presentar datos atípicos (observado en el boxplot) presentando un sesgo a derecha y colas pesadas.



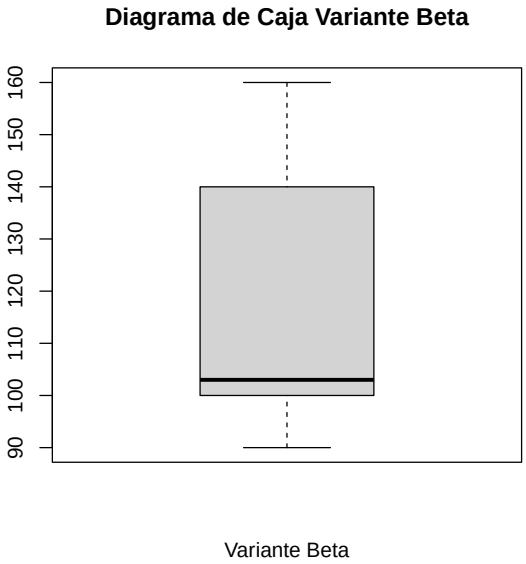
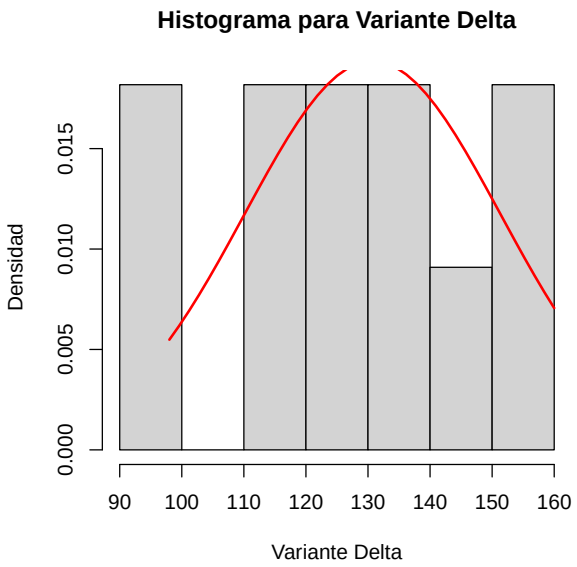


Tabla 1 : Descripción numérica para la variante beta

	Estimaciones
Promedio	118.89
Mediana	103.00
Varianza	690.86
Desvi.Estandar	26.28
Coef.Variación	22.11
Q1	100.00
Q2	103.00
Q3	140.00
Mínimo	90.00
Máximo	160.00

Análisis descriptivo Variante Delta

Es posible observar , aunque con pocos datos , la variante beta posee un comportamiento normal además de no presentar datos atípicos (observado en el boxplot) presentando simetría.



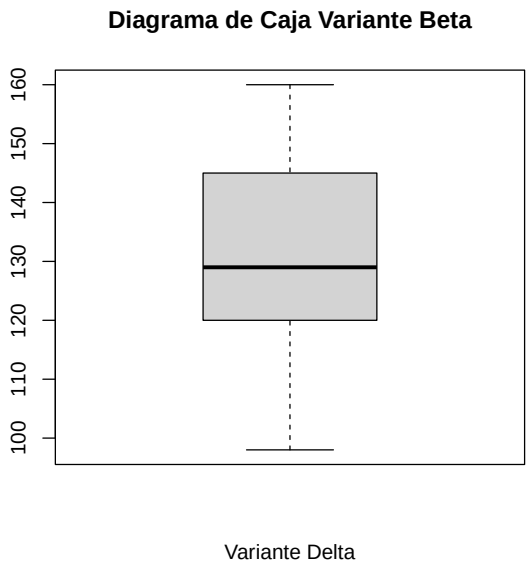


Tabla 2 : Descripción numérica para la Variante Delta

	Estimaciones
Promedio	130.73
Mediana	129.00
Varianza	424.82
Desvi.Estandar	20.61
Coef.Variación	15.77
Q1	120.00
Q2	129.00
Q3	145.00
Mínimo	98.00
Máximo	160.00

Del análisis descriptivo se observa una diferencia sustancial entre las medias y las varianzas de las personas infectadas con la variante beta y delta.

NOTA : siguiendo las recomendaciones del Dane para estimaciones e inferencias son los Criterios de precisión: Excelente, si su c.v.e. es menor del 3% De buena calidad, entre el 3% y el 5%; Aceptable, entre el 5% y el 15%; De uso restringido, si es mayor del 15%; estas estimaciones deben usarse con precaución.Por lo anterior quizá una inferencia con la media sobre los anteriores grupos (al ser mayores a 15%) deben tratarse con precaución.

Test de Normalidad

El test de Shapiro-Wilks plantea la hipótesis nula que una muestra proviene de una distribución normal. Se elige un nivel de significancia, por ejemplo 0,05, y tenemos una hipótesis alternativa que sostiene que la distribución no es normal.

Así:

$$H_0 : X \sim N(\mu, \sigma^2)$$
$$H_1 : X \nsim N(\mu, \sigma^2)$$

Ahora el test Shapiro-Wilks intenta rechazar la hipotesis nula al nivel de significancia y puesto que el tamaño de la muestra para ambos grupos es de menos de 50 individuos se usará el test de normalidad de Shapiro-Wilks al ser el mas potente.

```
##
##  Shapiro-Wilk normality test
##
## data:  vBeta$VBeta
## W = 0.87727, p-value = 0.1469
```

```
##
##  Shapiro-Wilk normality test
##
## data:  vDelta$VDelta
## W = 0.95095, p-value = 0.6561
```

dado los p-value de las pruebas siendo 0.1469 para el grupo con la variante Beta y 0.6561 el grupo con la variante Delta , con una significancia de 0.05 es posible decir que no existe evidencia para rechazar la hipótesis nula de normalidad para cada grupo . Es decir con una confianza del 95% los grupos se distribuyen aproximadamente normal.

Test de varianzas

Al obtener la normalidad de los grupos es posible realizar la prueba F de comparación de varianzas para grupo normales

```
##
##  F test to compare two variances
##
## data:  vBeta$VBeta and vDelta$VDelta
## F = 1.6263, num df = 8, denom df = 10, p-value = 0.4635
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.421867 6.984956
## sample estimates:
## ratio of variances
##           1.626251
```

El valor p de la prueba es de 0.4653, con una significancia de 0.05 no hay evidencia para rechazar la hipótesis nula de la igualdad de varianzas , es decir con una confianza del 95% los grupos provienen de poblaciones con varianzas iguales.

Test de Comparación de Medias con Varianzas iguales pero desconocidas

Definamos las variables X_1 =“cantidad de carga viral en personas infectadas con la variante Beta” y X_1 =“cantidad de carga viral en personas infectadas con la variante Beta” entonces sabemos por los puntos anteriores que $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$ y además que $\sigma_1^2 = \sigma_2^2$ pero desconocidas , así se plantea el contraste de hipótesis:

$$H_0 : \mu_1 - \mu_2 = 0 \quad VS \quad H_1 : \mu_1 - \mu_2 \neq 0$$

para hallar una diferencia significativa entre los grupos de personas infectadas , veamos el estadístico de prueba

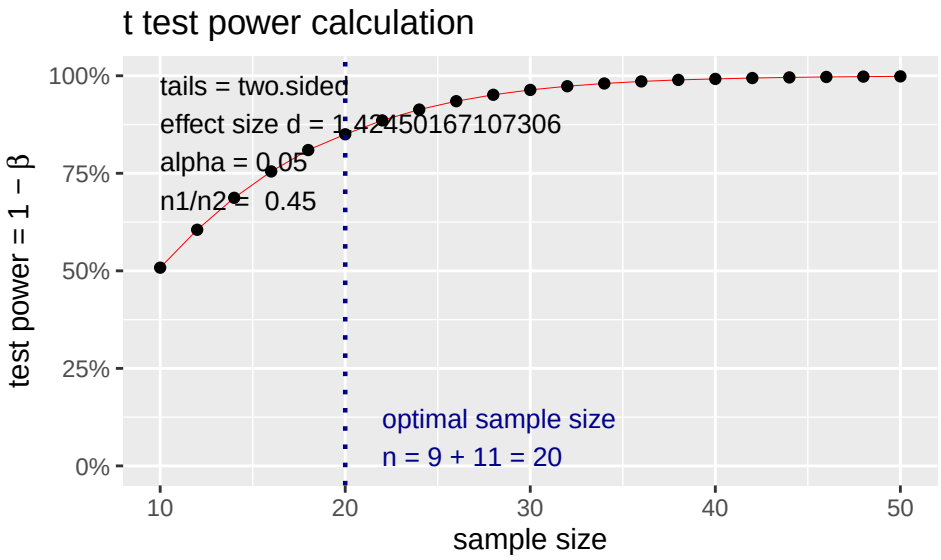
$$t_c = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{S_p\sqrt{1/n_1 + 1/n_2}}$$

donde $S_p = \frac{(n_1-1)s_1^2+(n_1-1)s_2^2}{n_1+n_2-2}$ al tener esto en cuenta se usará la prueba para comparación en R

```
##
##  Two Sample t-test
##
## data:  vBeta$VBeta and vDelta$VDelta
## t = -1.1302, df = 18, p-value = 0.2732
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -33.84387  10.16711
## sample estimates:
## mean of x mean of y
##  118.8889  130.7273
```

puesto que el valor P es de 0.2732 con una significancia del 0.05 no hay evidencia para rechazar la hipótesis nula de la igualdad de medias ,por lo tanto con una confianza del 95% no existe evidencia significativa para decir que hay una cantidad media de partículas del virus en la variantes de cada grupo.

Poder estadístico y tamaño del efecto



El poder de la prueba realizada es de aproximadamente el 85% es decir la probabilidad de rechazar la hipotesis nula cuando esta es falasa.

El tamaño del efecto representa el grado en que la hipótesis nula es falsa.El efecto de la prueba de medias para grupos independientes es la de d de cohen calculada de la siguiente manera

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_{comun}} = \frac{|118.89 - 130.73|}{23.3010} = 0.5081$$

donde $S_{comun} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(9-1)26.28^2+(11-1)20.61^2}{9+11-2}} = 23.3010$ usando la tabla de la figura 1. podemos decir que el tamaño del efecto es medio

Table 1 Values of Effect Sizes and Their Interpretation

Kind of Effect Size	*Small*	*Medium*	*Large*
r	.10	.30	.50
d	0.20	0.50	0.80
η^2_p	.01	.06	.14
f^2	.02	.15	.35

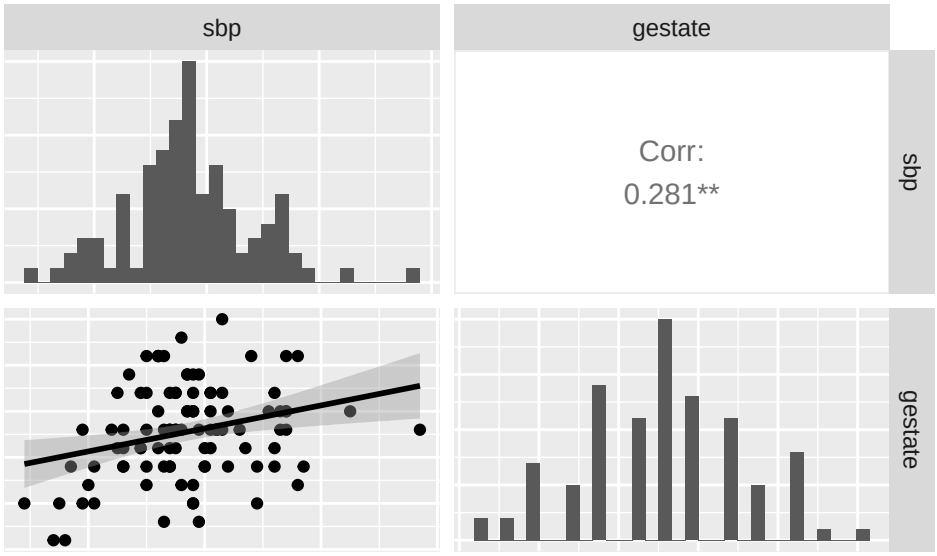
Source: Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155

Figure 1: cohen-tabla

Regresión Lineal

El conjunto de datos lowbwt contiene información sobre 100 niños nacidos con bajo peso en Boston. Las medidas de presión sistólica están en la columna sbp y los valores de tiempo gestacional en gestage

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



En el diagrama de dispersión no se observa ninguna relación lineal aparente además de obtener un valor de la correlación bajo.

realizamos el modelo de regresión lineal de la forma

$$(sbp)_i = \beta_0 + \beta_1(gestate)_i + e_i$$

```
##
## Call:
## lm(formula = sbp ~ gestate, data = pesoBajo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.162  -7.828  -1.483   5.568  39.781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5521    12.6506   0.834  0.40625
## gestate       1.2644     0.4362   2.898  0.00463 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11 on 98 degrees of freedom
## Multiple R-squared:  0.07895,    Adjusted R-squared:  0.06956
## F-statistic: 8.401 on 1 and 98 DF,  p-value: 0.004628
```

Se puede observar que el valor p para la prueba para el intercepto es de 0.40625 por lo tanto el intercepto no es significativo para el modelo.Por lo tanto se propone el modelo sin intercepto

$$(sbp)_i = \beta_1(gestate)_i + e_i$$

```
##
## Call:
## lm(formula = sbp ~ gestate + 0, data = pesoBajo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.672  -7.925  -1.366   5.884  39.821
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## gestate  1.62687    0.03787   42.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.98 on 99 degrees of freedom
## Multiple R-squared:  0.9491, Adjusted R-squared:  0.9486
## F-statistic:  1845 on 1 and 99 DF,  p-value: < 2.2e-16
```

es posible observar que el valor p para la pendiente del modelo es aproximadamente 0 es decir la pendiente (β_1) es significativa para el modelo presentado ,además, se tiene un valor $R^2 = 0.9491$ así el modelo es capaz de explicar un 94.91% de variación de la presión sistólica. Por lo tanto el modelo de regresión es un buen modelo lineal para explicar la relación entre la presión sistólica y el tiempo gestacional de los recien nacidos.

El modelo ajustado es $(\hat{sbp})_i = 1.62687(gestate)_i$, esta pendiente representa que por cada semana de gestación la presión sisólica aumenta en aproximadamente 1.62687unidades.

Cuando se tiene una observación $X_0 = 31$ semanas la estimación puntual del modelo para este valor es $31(1.62687) = 50.43297$ Unidades.

Comprobación de supuestos

Normalidad en los residuos

puesto que n=100 procederemos a usar el test de kolmogorov-smirnov para probar ls normalidad de los residuos .

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuos
## D = 0.082948, p-value = 0.0865
```

para $\alpha = 0.05$, no se rechaza la hipótesis nula de que la distribución de los residuales es normal.

Homocedasticidad de los residuos

Se usará el test Goldfeld-Quandt pues el test Brush-pagan exige que nuestro modelo cuente con pendiente.

```
##
##  Goldfeld-Quandt test
##
## data:  regresion2
## GQ = 0.56683, df1 = 49, df2 = 49, p-value = 0.9752
## alternative hypothesis: variance increases from segment 1 to 2
```

el valor p es de 0.9752 no hay evidencia para rechazar H_0 por lo tanto la varianza no incrementa de un grupo a otro por lo tanto es homocedastica con una confianza del 95%

Independencia de los residuos

```
##
##  Durbin-Watson test
##
## data:  regresion2
## DW = 1.8103, p-value = 0.1692
## alternative hypothesis: true autocorrelation is greater than 0
```

con un valor p de 0.1692 no hay evidencia para rechazar la hipótesis nula con una significancia de 0.05 , por lo tanto los residuos son independientes con un nivel de confianza del 5%.

Conclusión

Puesto que el modelo cumple todos los supuesto y tenemos valor de R^2 alto , nuestro ajuste es bueno y el modelo sirve para realizar predicciones en el rango de la covariable o intervalos de confianza para una nueva observación.