Universidad Austral de Chile

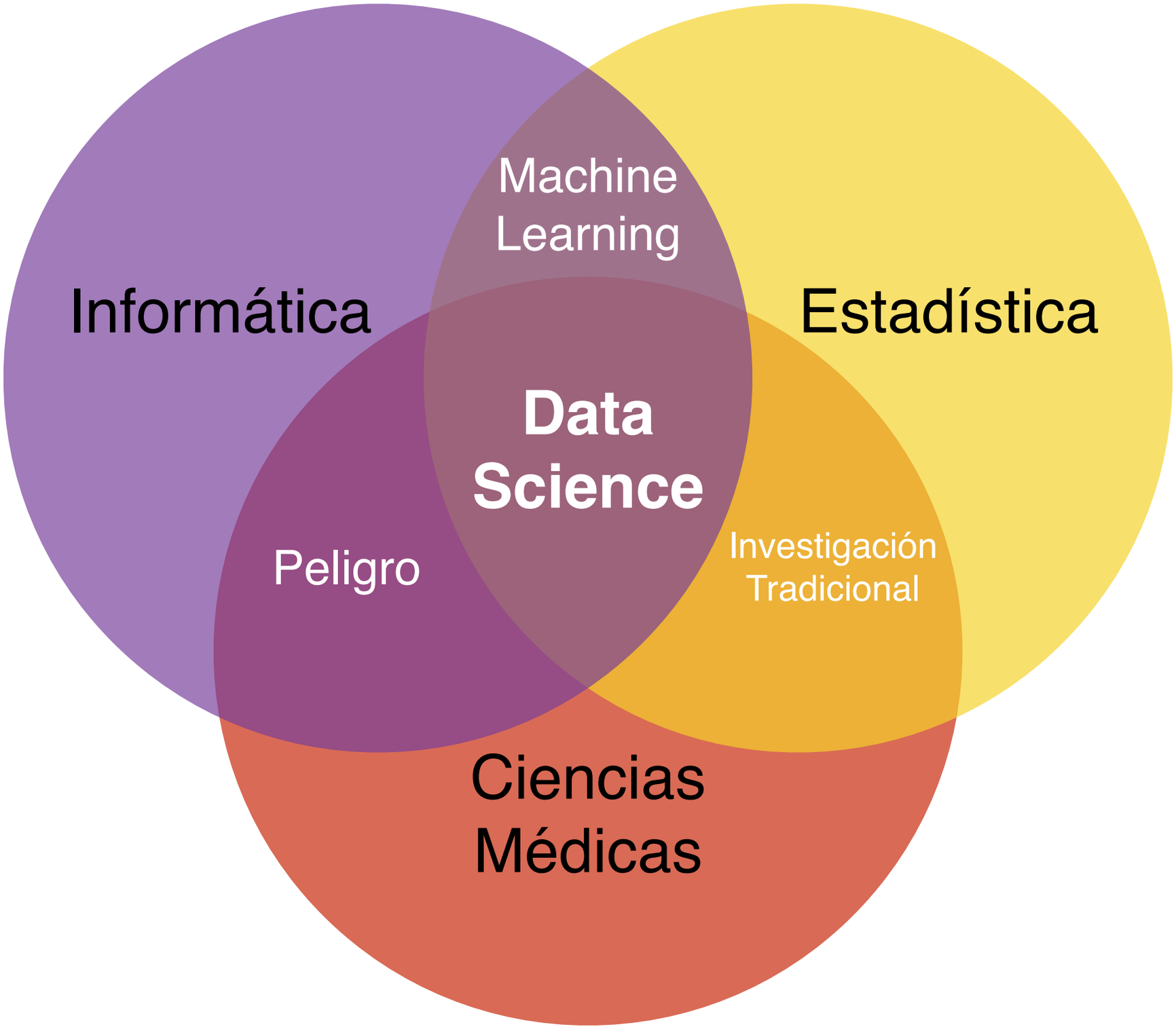
Conocimiento y Naturaleza

Herramientas estadísticas aplicada a la genómica

BIMI431 - Estadística y Genómica

Diego Halabi, DDS, PhD - Laboratorio de Cronobiología del Desarrollo - 24/09/2020

Manejo de Datos (Data Science)



Estadística

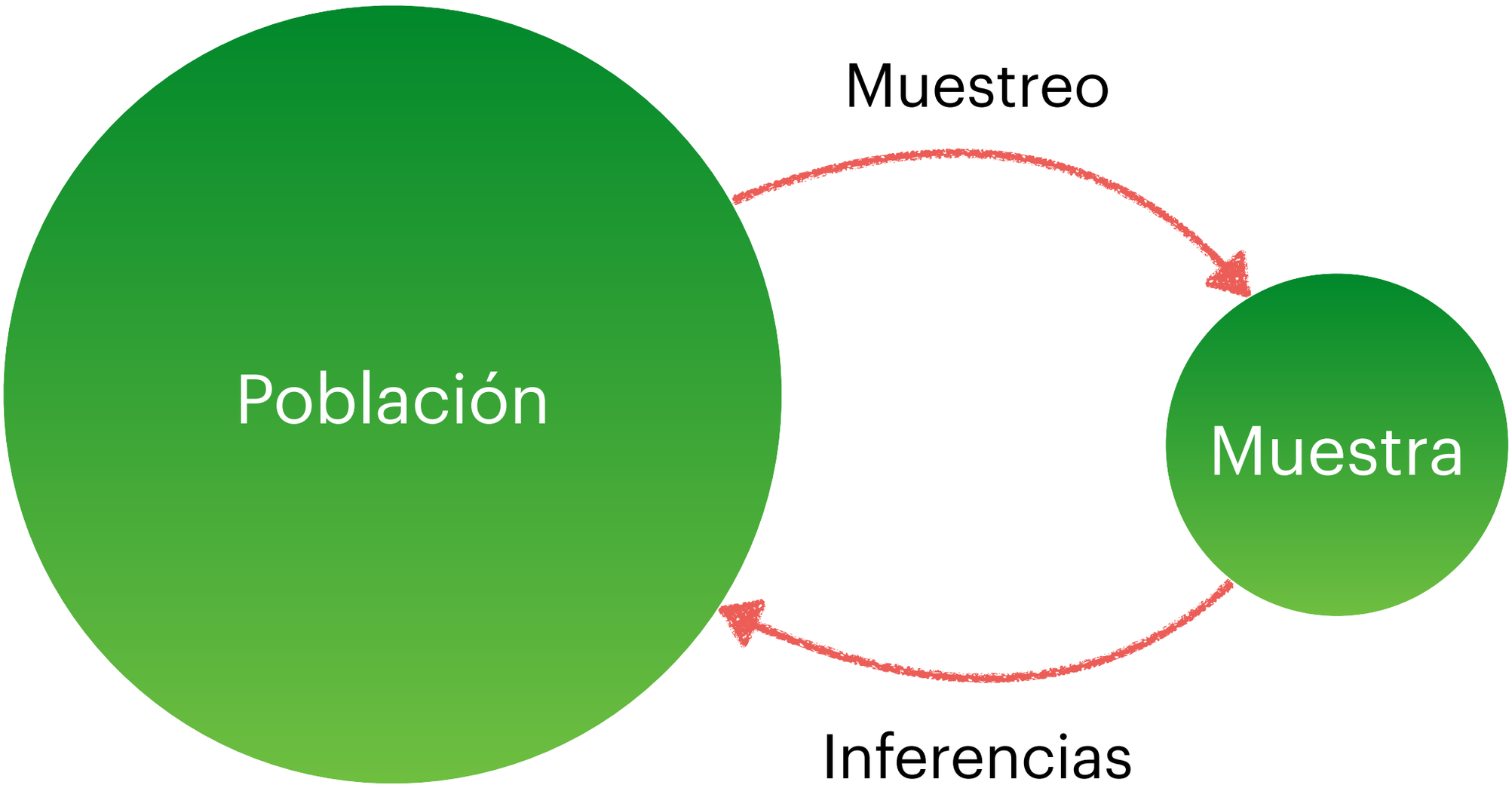
Describir

Resumir

Organizar

Inferir

Estadística en investigación



Variables

Cualquier elemento susceptible de ser medido.
También los denominaremos *vectores*.

Exhaustivas; ningún valor puede quedar fuera

Excluyentes; ningún valor puede ser incluido en 2 o más categorías

Tipos de variable

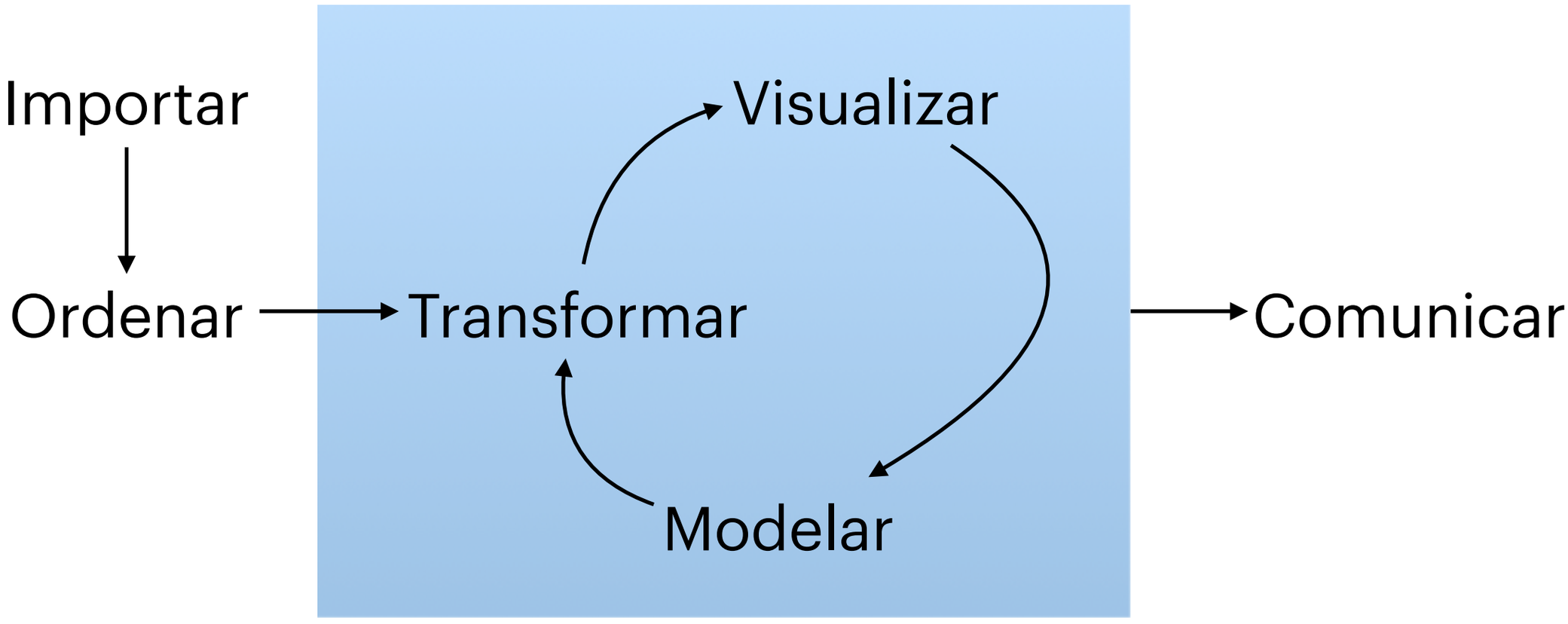
Categóricas

Variable	Ejemplo
Dicotómica (logical)	"True", "False"
Nominal (character)	"amarillo"
Factor	"Control", "Tratamiento 1", "Tratamiento 2"
Ordinal	Escala likert

Numéricas

Variable	Ejemplo
Discretas (integer)	100, 200, 300
Continuas (numeric)	1.5, 35.2, 4.03

Manejo de datos

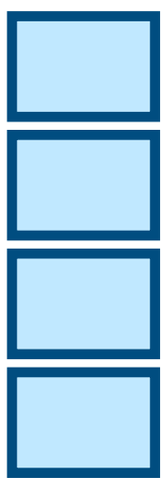


Recolección de datos

Importación y ordenamiento de los datos (tabulación)

Estructura de los datos

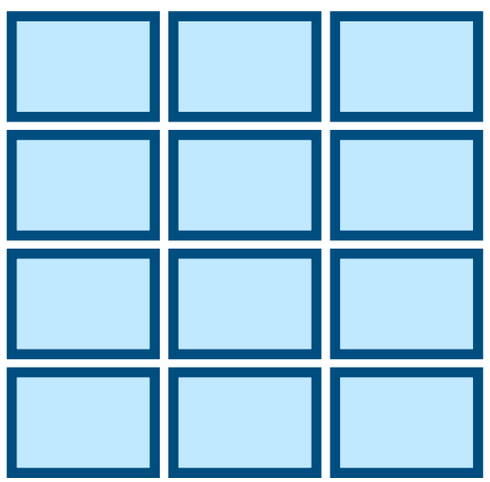
Vector



1 columna de datos

1 tipo de variable

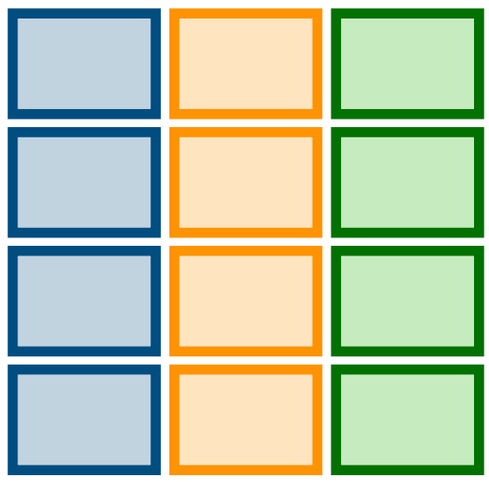
Matriz



múltiples columnas de datos

1 tipo de variable

Data frame



múltiples columnas de datos

múltiples tipos de variable

Tabulación

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	6.67	32	1	0
A	8.43	40	2	0
B	6.01	31	0	0
A	7.78	35	0	0
A	7.89	36	1	1
B	6.41	31	2	0
B	8.90	41	0	0
B	5.56	30	-	0
A	7.33	33	1	1
A	8.21	39	0	0
B	7.09	34	0	0
B	6.34	31	1	1
A	8.17	42	2	1

Visualización de los datos

Estadística descriptiva

Resumir } Medidas de tendencia central

Describir } Medidas de dispersión

Presentar } Tablas y gráficos

Estadística descriptiva

Se presenta la información como números o gráficos

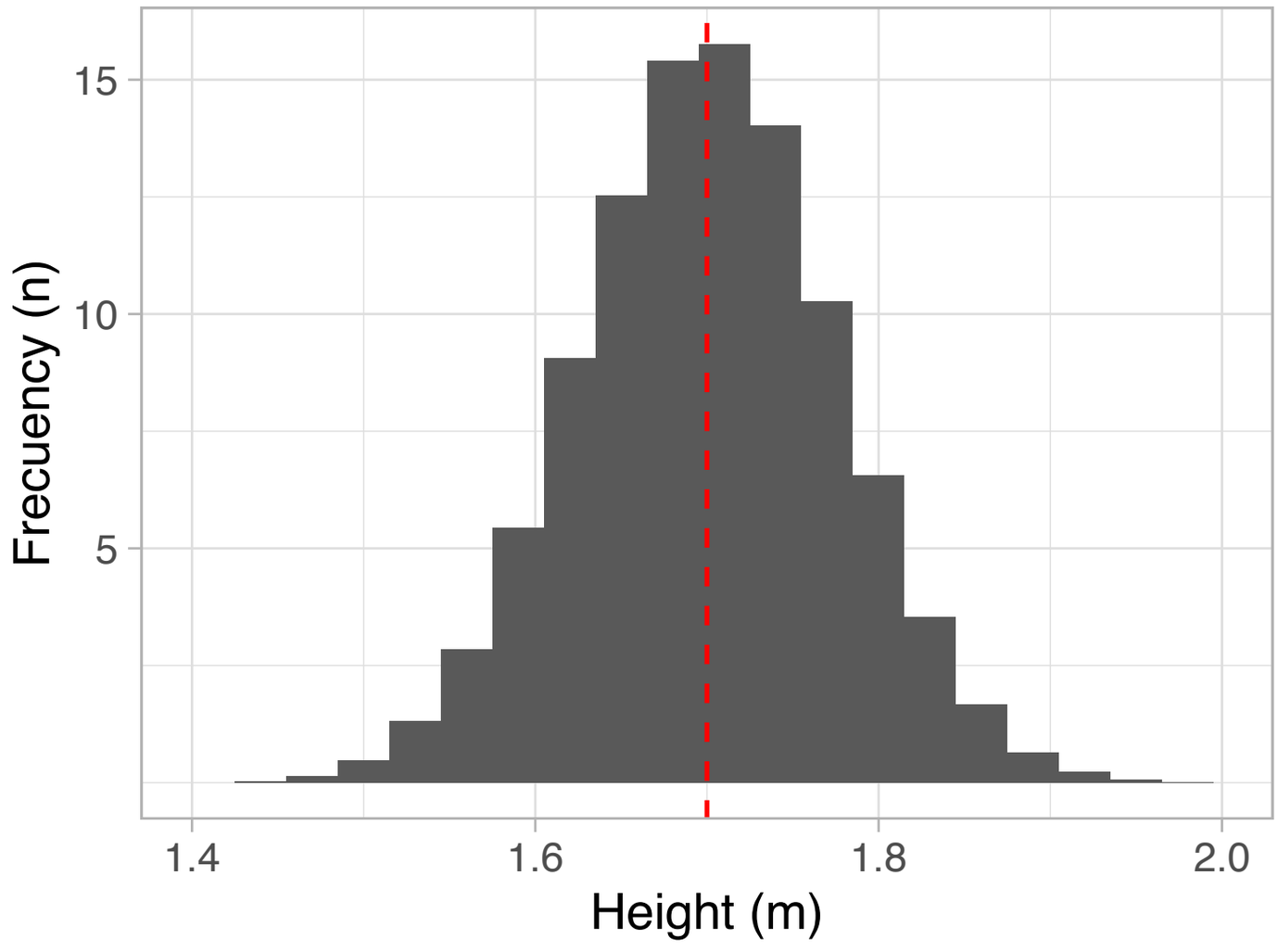
Números: tablas

Figuras: Gráficos

Números

Medidas de **tendencia central**; media, mediana, moda

Medidas de **dispersión**; desviación estándar, varianza, rango, recorrido intercuartil



Inferencia estadística

Hipótesis nula

Sentencia afirmativa, **cuantificable**; diferencia de medias, tasas, etc.

Es aceptable que no esté implícita en el texto; puede deducirse del objetivo.

Ejemplo:

“No hay diferencias estadísticamente significativas en el valor medio de HbA1c entre tratamiento A v/s tratamiento B”.

Hipótesis

↓

Metodología

↓

Resultados

↓

Realidad

Verdad

Error aleatorio

Error sistemático

Hipótesis

↓

Metodología

↓

Resultados

↓

Realidad

Verdad

Resultados

Verdad	No diferencias (acepto H_0)	Diferencias (rechazo H_0)
No diferencias	No error	Error tipo I
Diferencias	Error tipo II	No error

Nivel de confianza

α (Nivel de significancia)

Resultados

Verdad	No diferencias (acepto H_0)	Diferencias (rechazo H_0)
No diferencias	No error 1 - α	Error tipo I
Diferencias	Error tipo II	No error 1 - β

β

Potencia

Contraste de hipótesis

Test estadístico

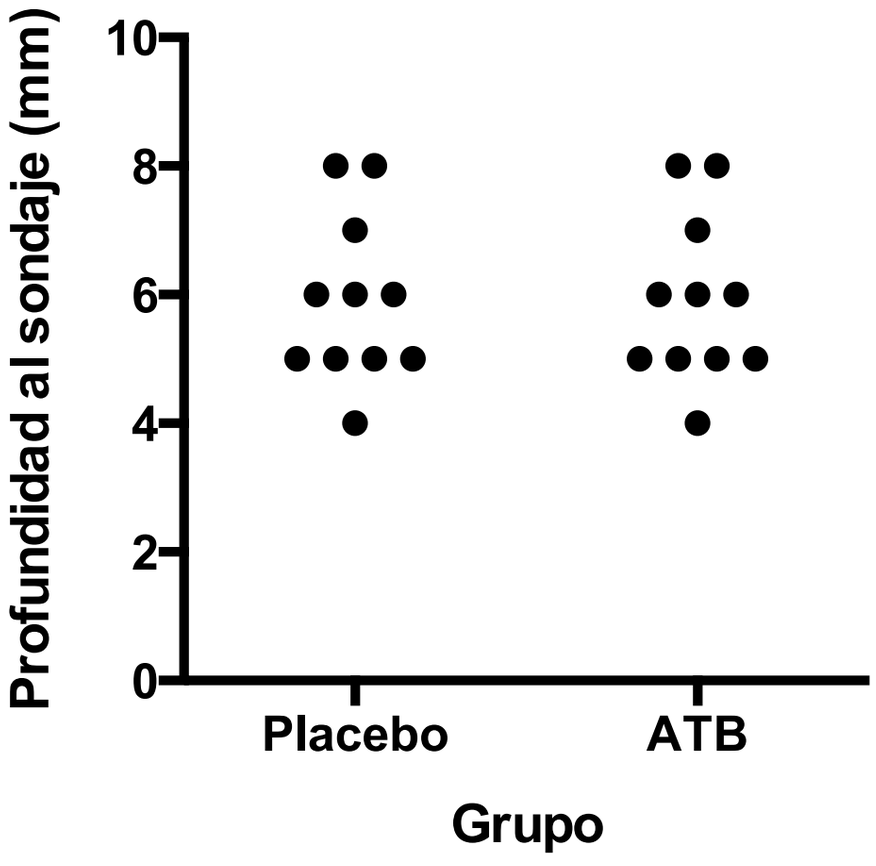
Prueba de significación

Análisis estadístico

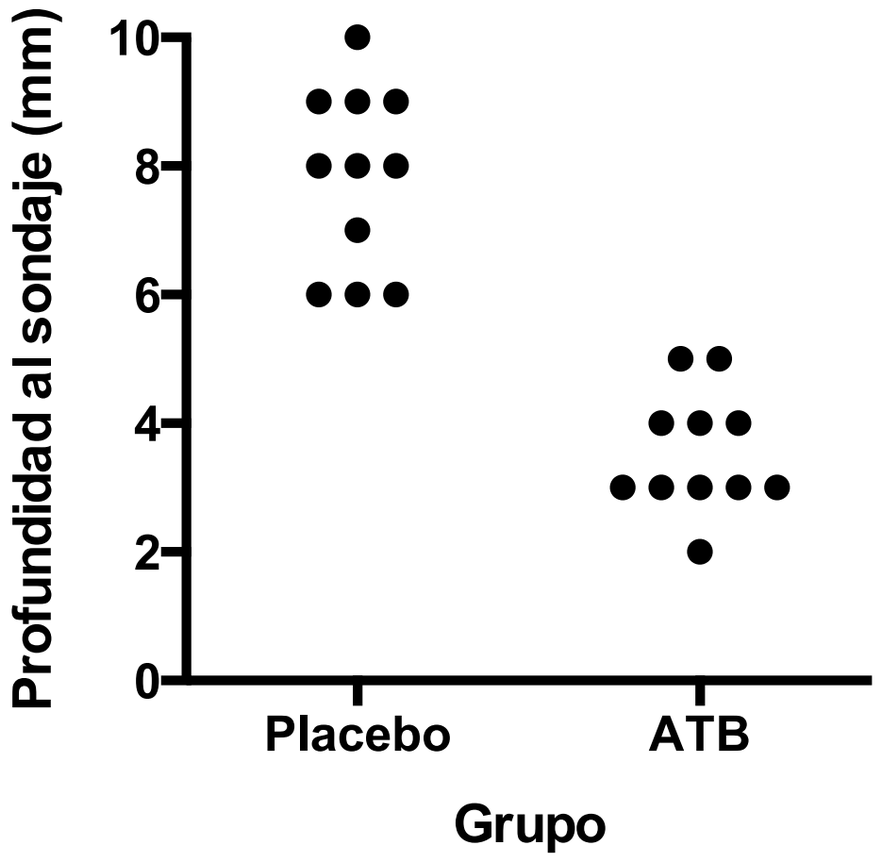
¿p > 0.05?

Contraste de hipótesis

Hipótesis nula

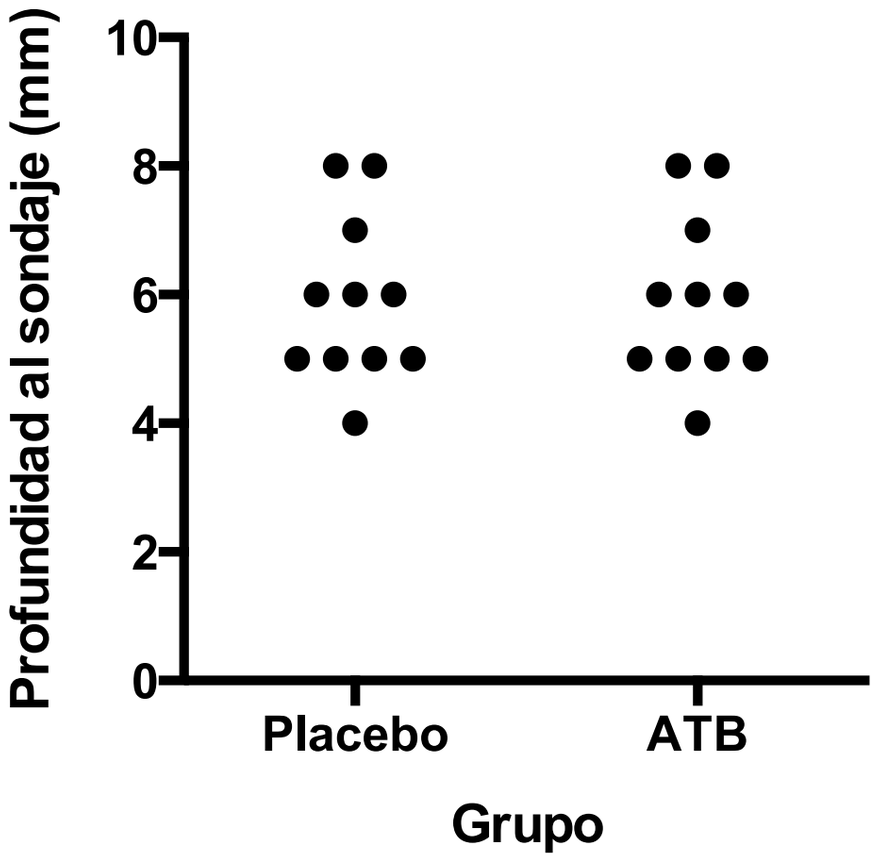


Resultados estudio

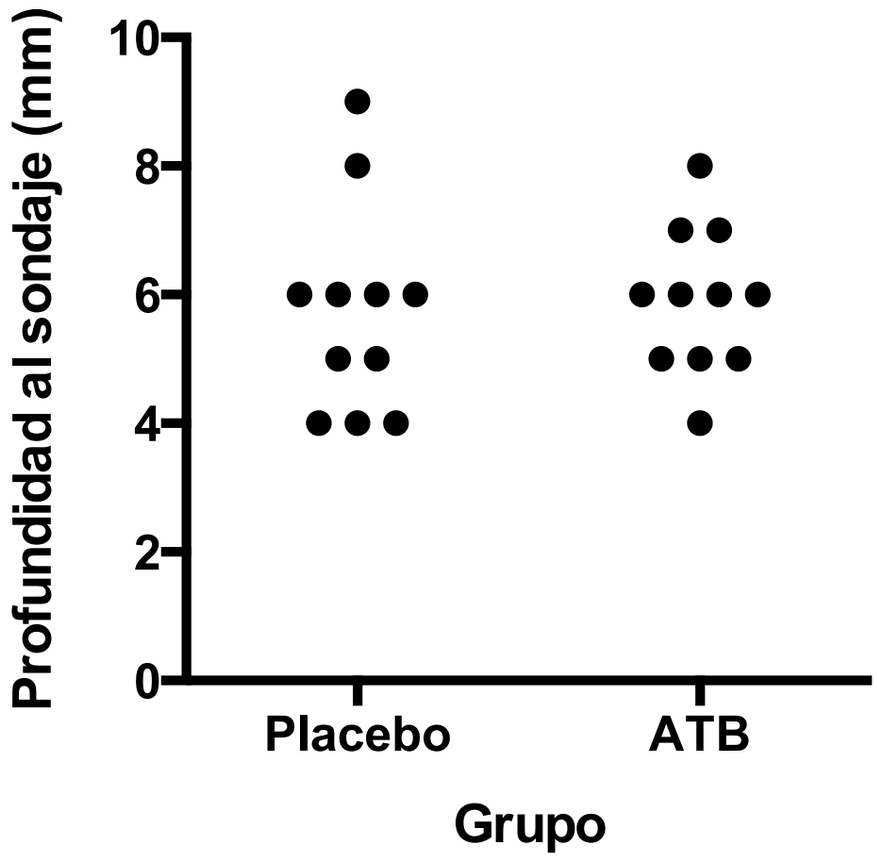


Contraste de hipótesis

Hipótesis nula

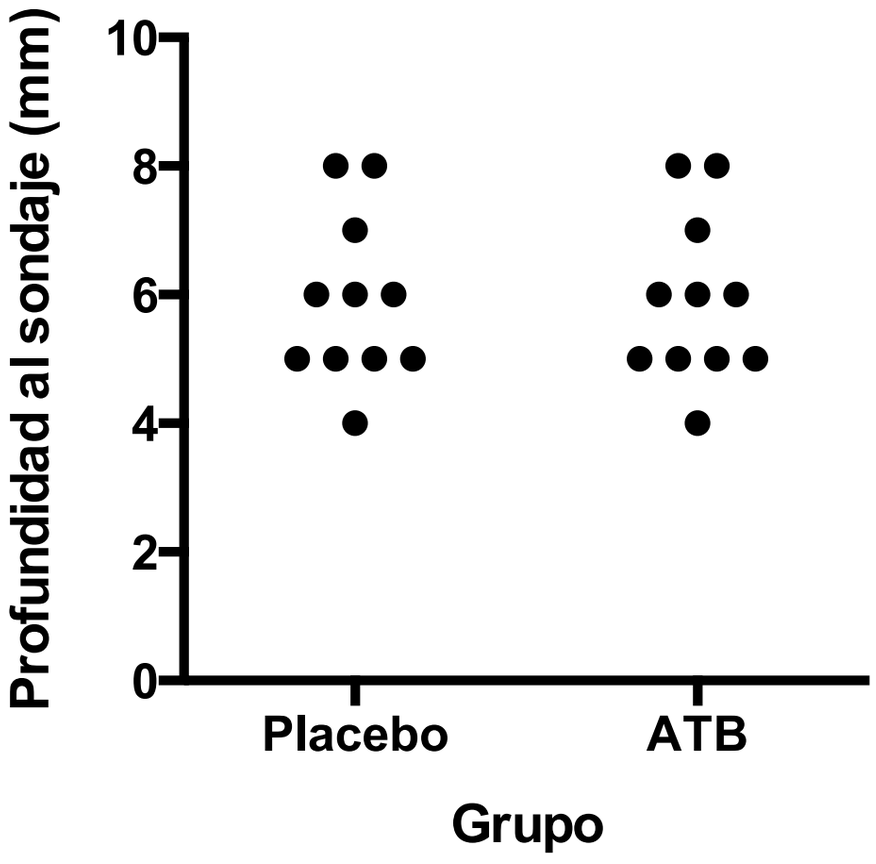


Resultados estudio

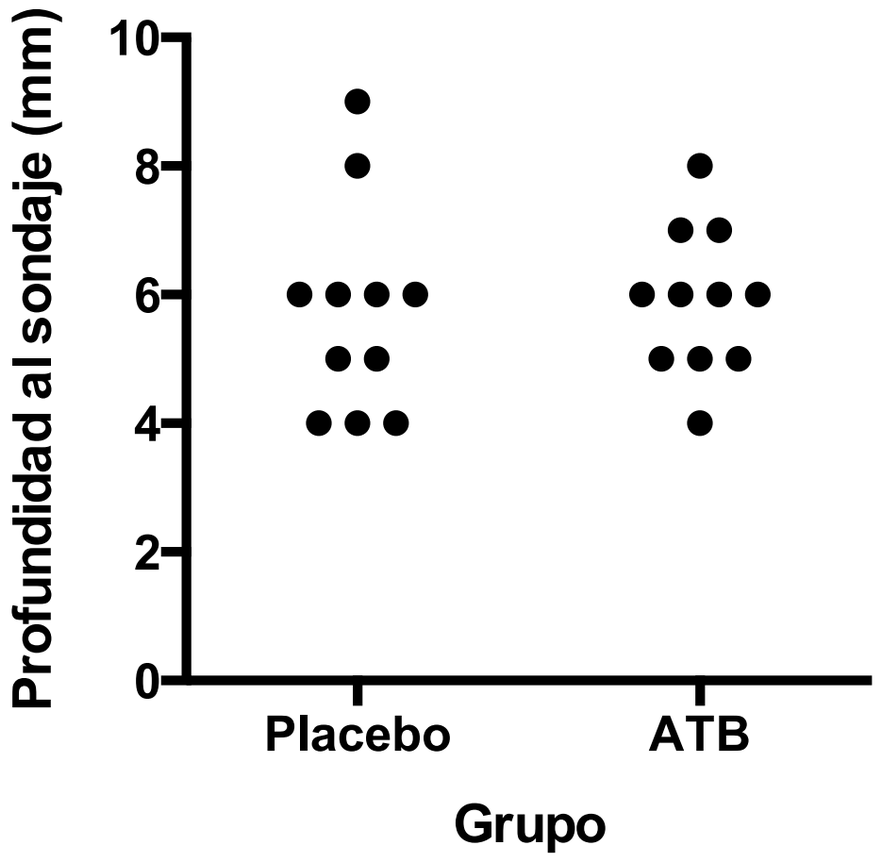


Contraste de hipótesis

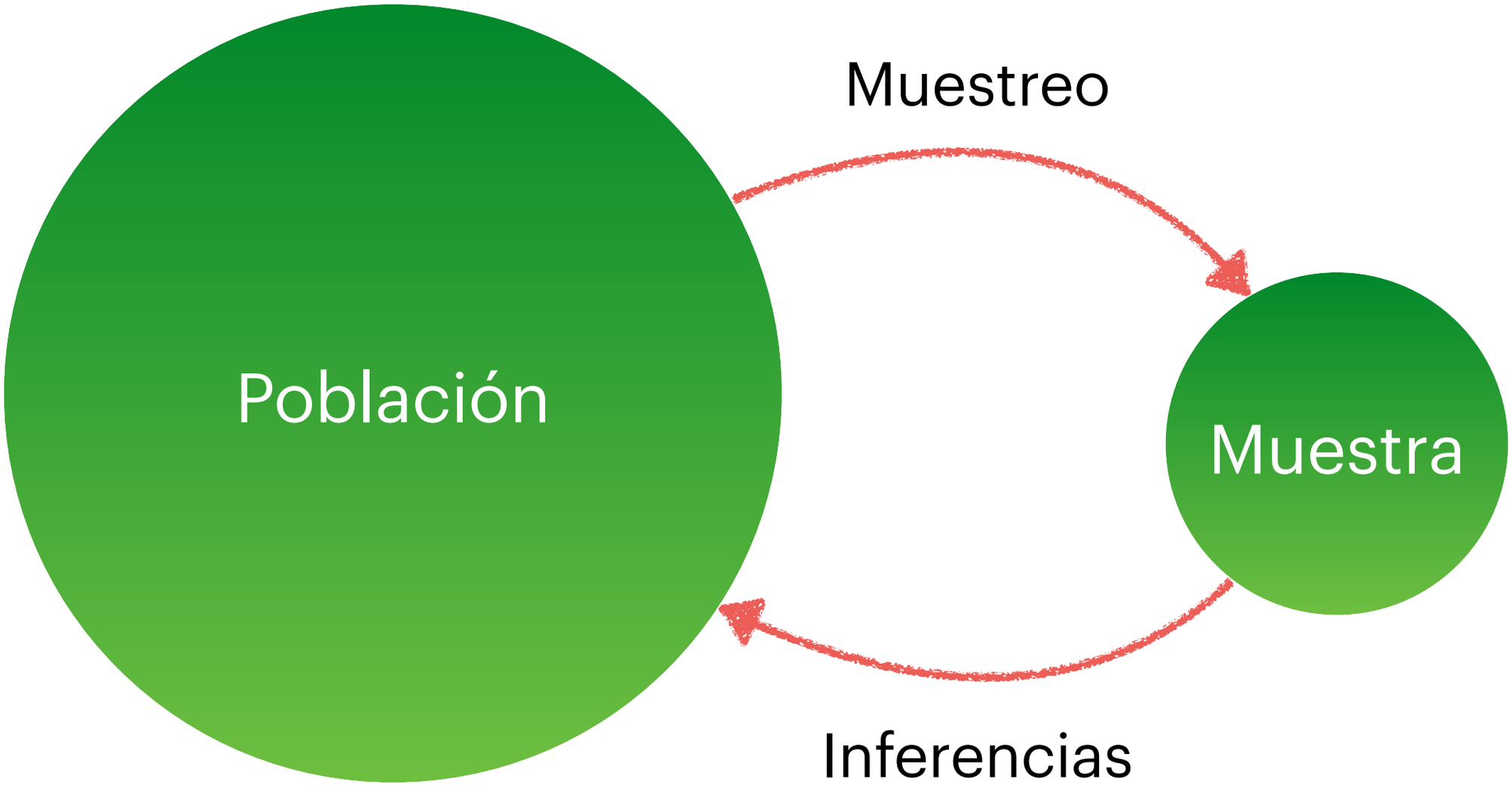
Hipótesis nula



Resultados estudio



α → **p-value**


Muestreo
Población
Muestra
Inferencias

Elección del test estadístico

Input data

Define α

Normality test

Shapiro - Wilk

p < α

p > α

2 group

> 2 group

2 group

> 2 group

paired

unpaired

Kruskal - Wallis

t-test

p > α

Levene

p < α

Games - Howell

Wilcoxon

U Mann - Witney

1 factor

2 factor

Anova

2-way Anova

Significant?

p > α

p < α

Nothing significant

Post - hoc*

*Only for > 2 group tests

Power

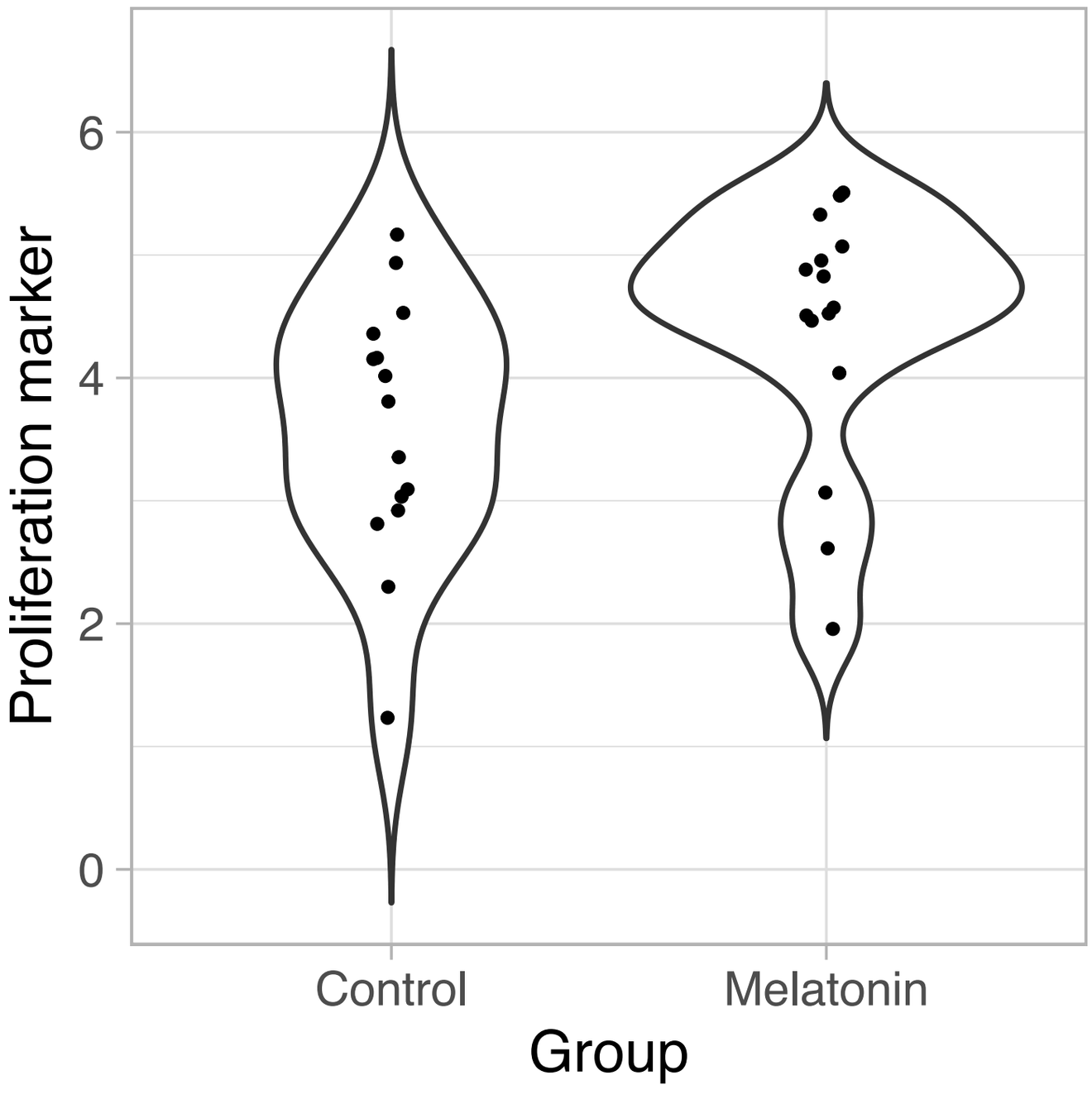
Communicate

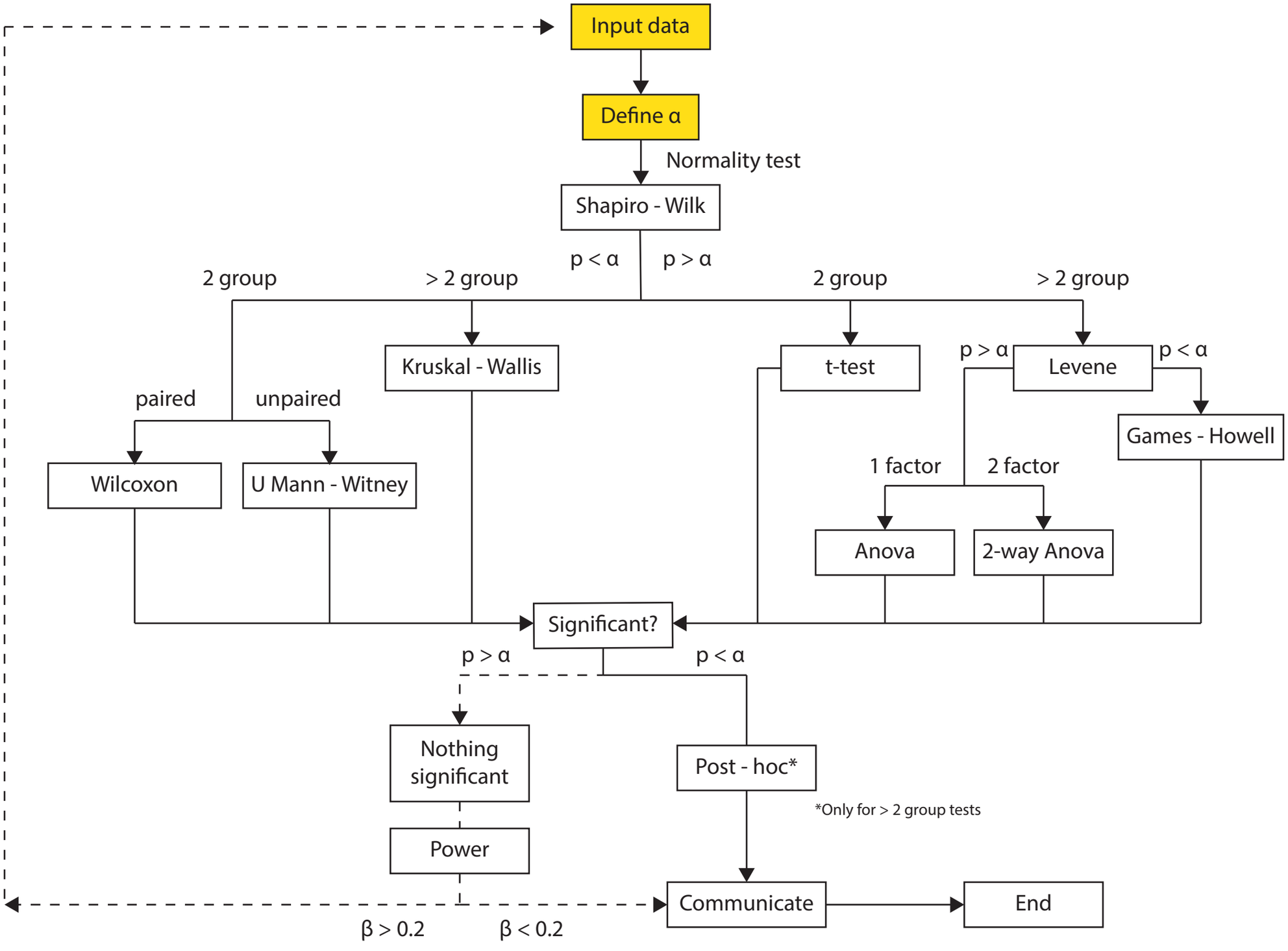
End

β > 0.2

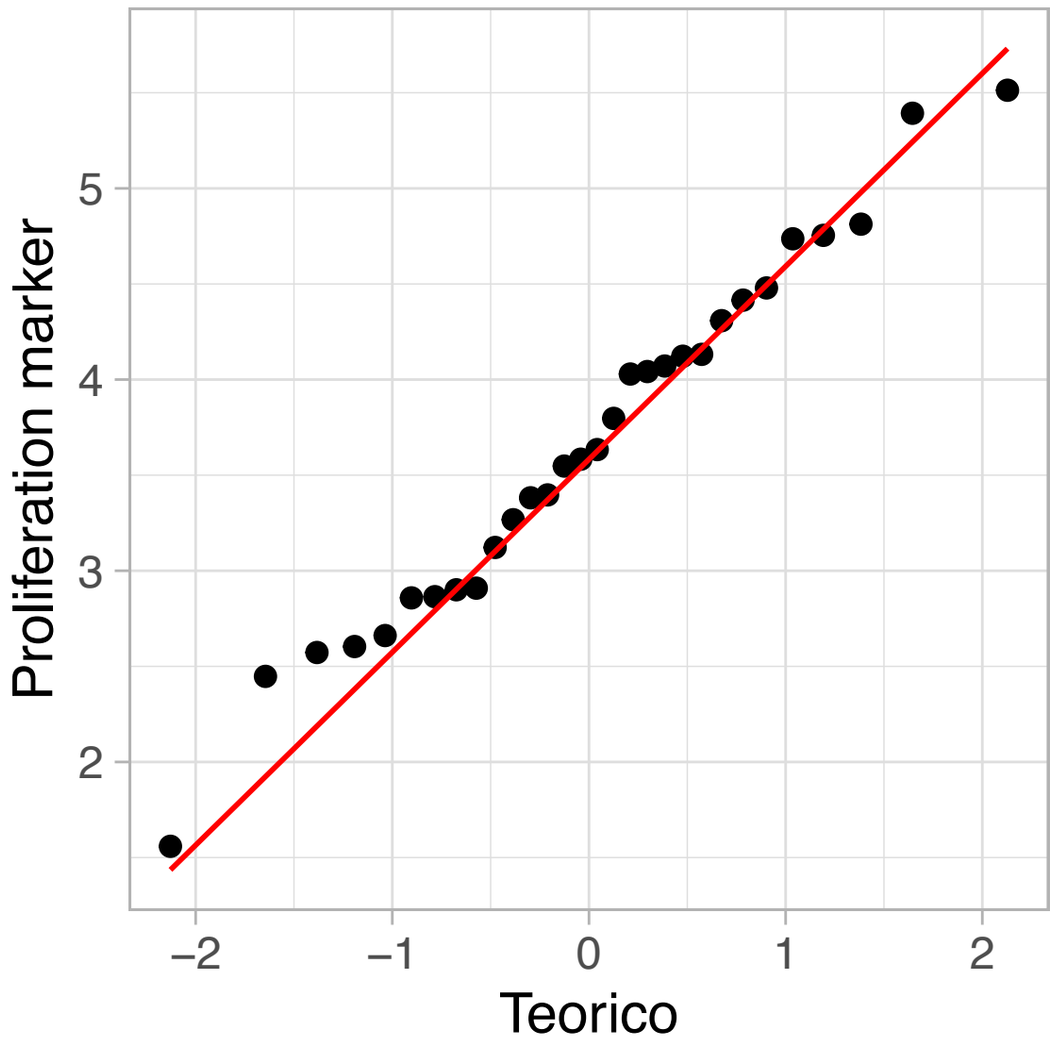
β < 0.2

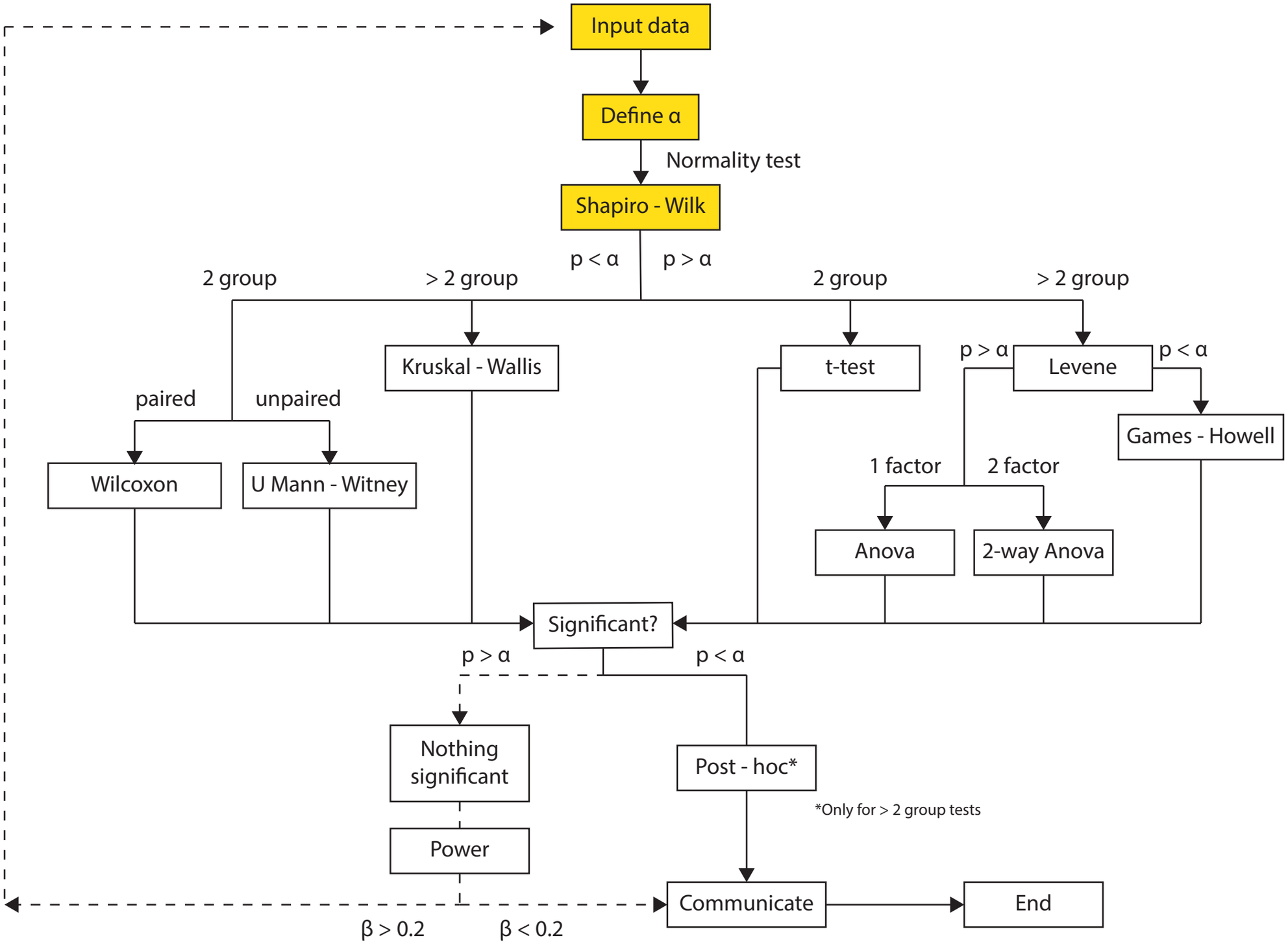
```
> str(df)
'data.frame':   30 obs. of  2 variables:
 $ Group        : Factor w/ 2 levels "Control","Melatonin": 1 1 1 1 1 1 1 1 1 1 ...
 $ Proliferation: num  2.3 3.35 4.16 2.81 3.09 ...
```



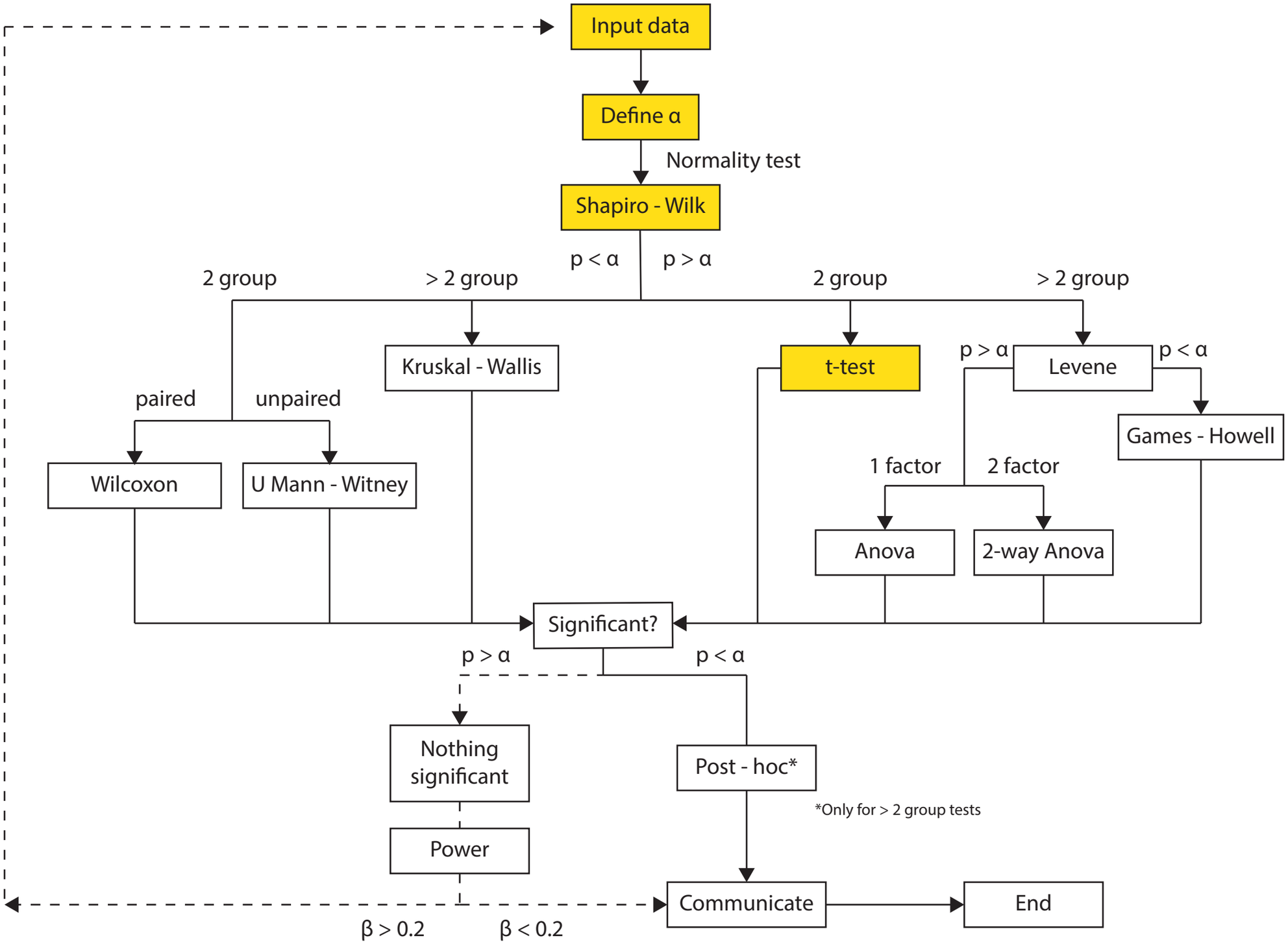

Input data
Define α
Normality test
Shapiro - Wilk
p < α
p > α
2 group
> 2 group
2 group
> 2 group
paired
unpaired
Kruskal - Wallis
t-test
p > α
Levene
p < α
Games - Howell
1 factor
2 factor
Wilcoxon
U Mann - Witney
Anova
2-way Anova
Significant?
p > α
p < α
Nothing significant
Post - hoc*
*Only for > 2 group tests
Power
Communicate
End
β > 0.2
β < 0.2

```
> shapiro.test(df$Proliferation)

    Shapiro-Wilk normality test

data:  df$Proliferation
W = 0.9838, p-value = 0.915
```

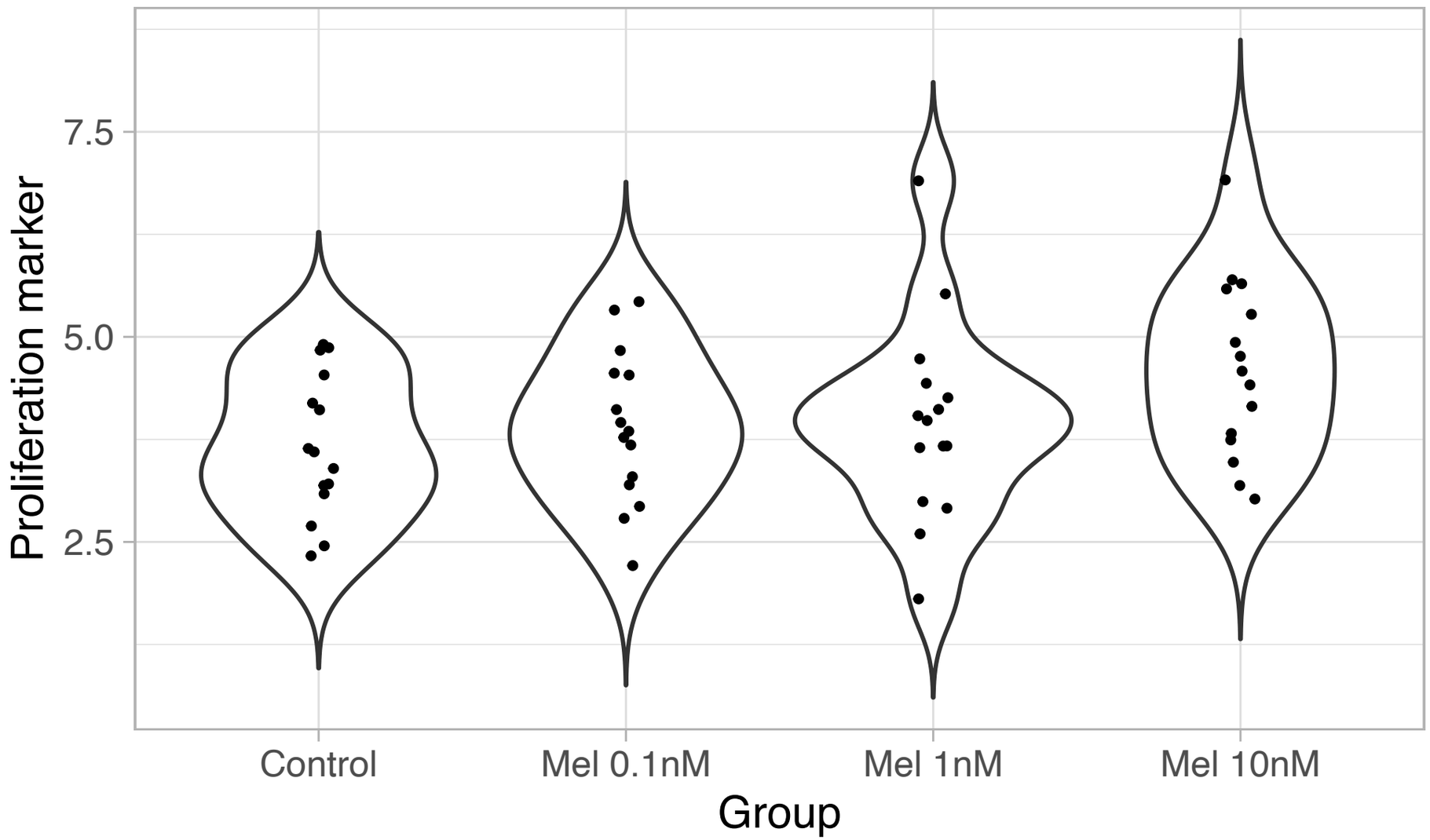


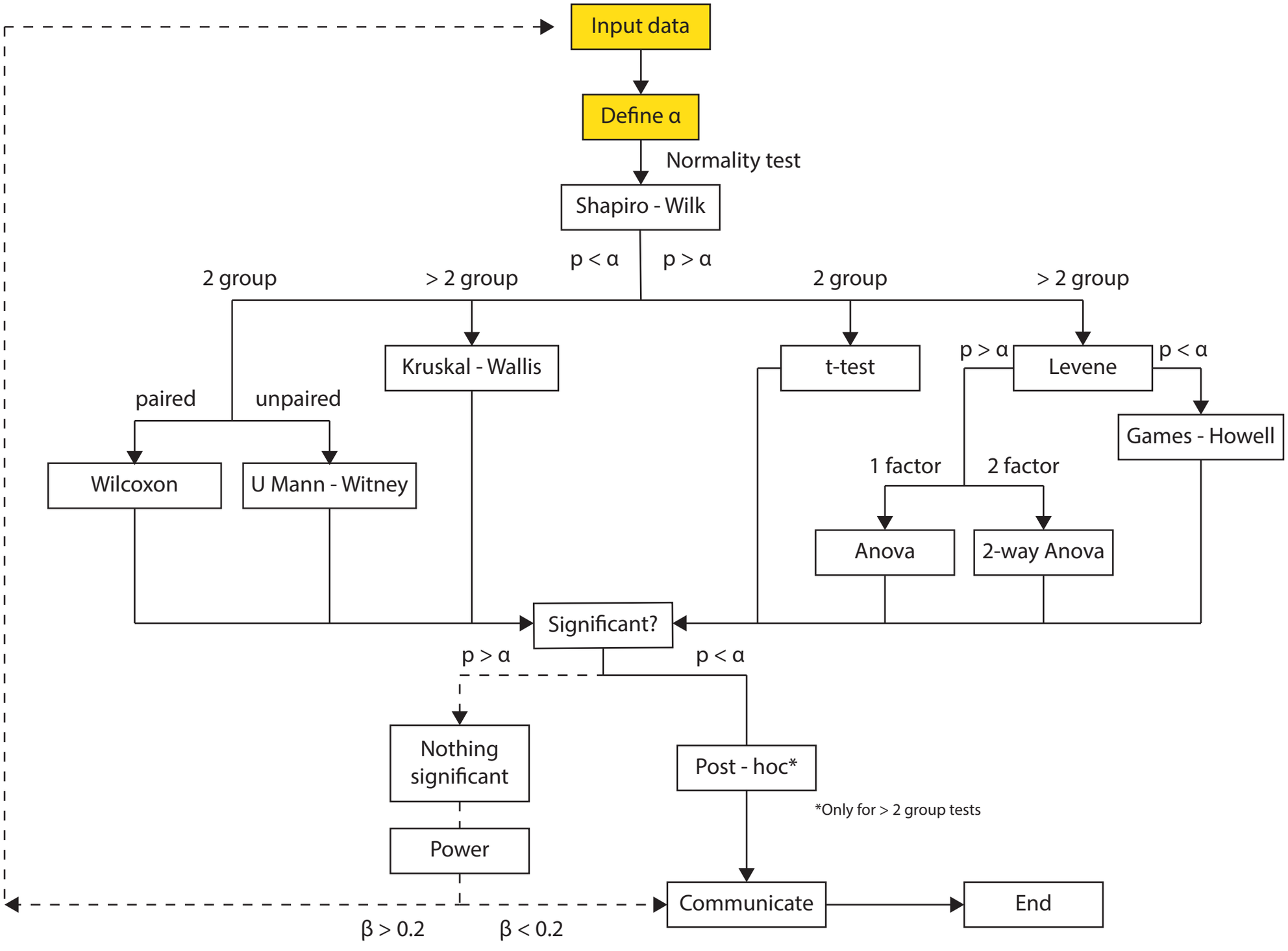

Input data
Define α
Normality test
Shapiro - Wilk
p < α
p > α
2 group
> 2 group
2 group
> 2 group
paired
unpaired
Kruskal - Wallis
t-test
p > α
Levene
p < α
Games - Howell
1 factor
2 factor
Wilcoxon
U Mann - Witney
Anova
2-way Anova
Significant?
p > α
p < α
Nothing significant
Post - hoc*
*Only for > 2 group tests
Power
Communicate
End
β > 0.2
β < 0.2

```
> t.test(df$Proliferation~df$Group, var.eq = T)

    Two Sample t-test

data:  df$Proliferation by df$Group
t = -2.9108, df = 28, p-value = 0.006995
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4960887 -0.2601717
sample estimates:
  mean in group Control mean in group Melatonin
               3.224674                4.102805
```


Input data
Define α
Normality test
Shapiro - Wilk
p < α
p > α
2 group
> 2 group
2 group
> 2 group
paired
unpaired
Kruskal - Wallis
t-test
p > α
Levene
p < α
Games - Howell
1 factor
2 factor
Wilcoxon
U Mann - Witney
Anova
2-way Anova
Significant?
p > α
p < α
Nothing significant
Post - hoc*
*Only for > 2 group tests
Power
Communicate
End
β > 0.2
β < 0.2

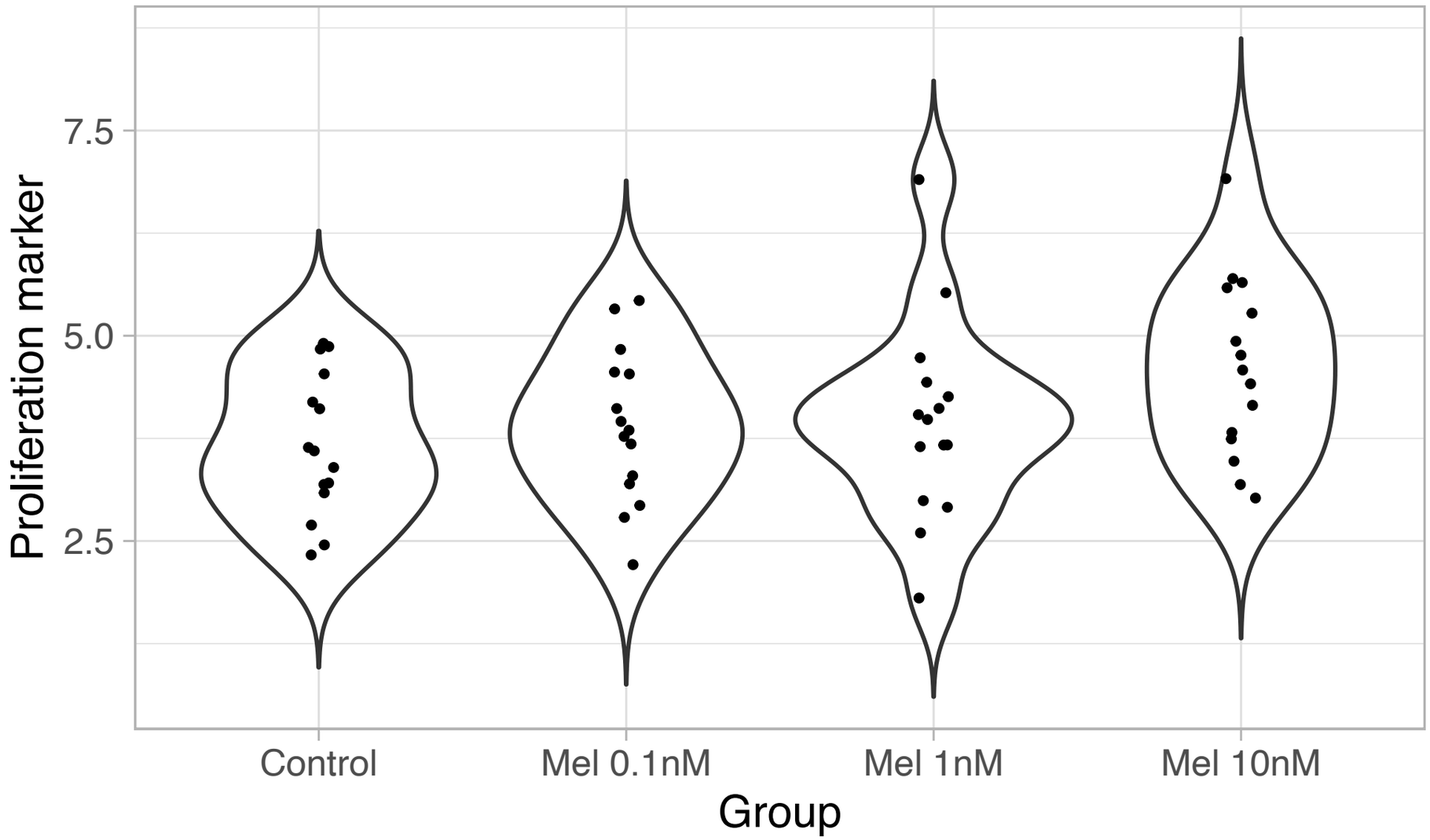
```
'data.frame':   60 obs. of  2 variables:
 $ Group        : Factor w/ 4 levels "Control","Mel 0.1nM",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Proliferation: num  3.21 3.09 2.45 4.19 3.6 ...
```

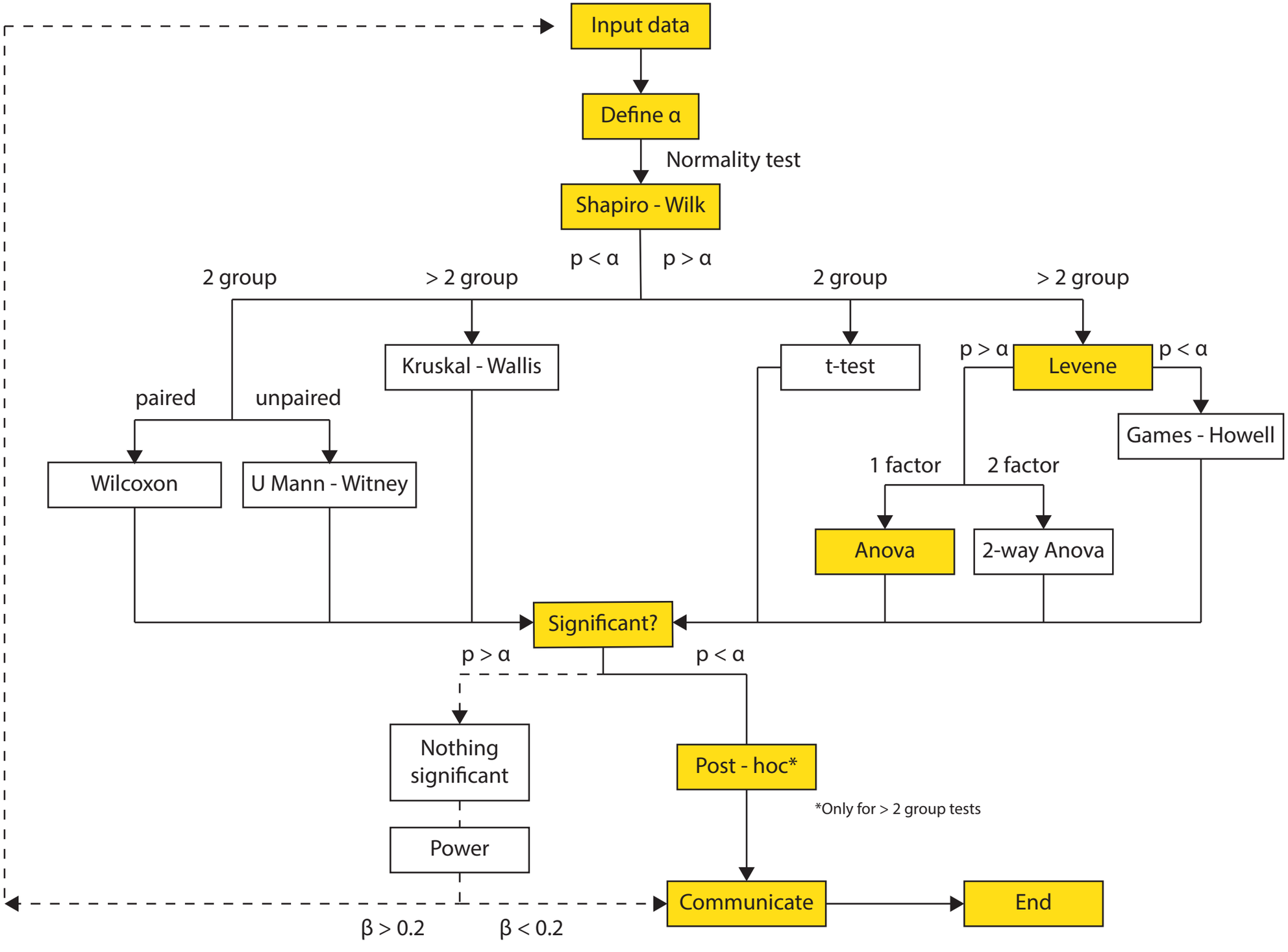


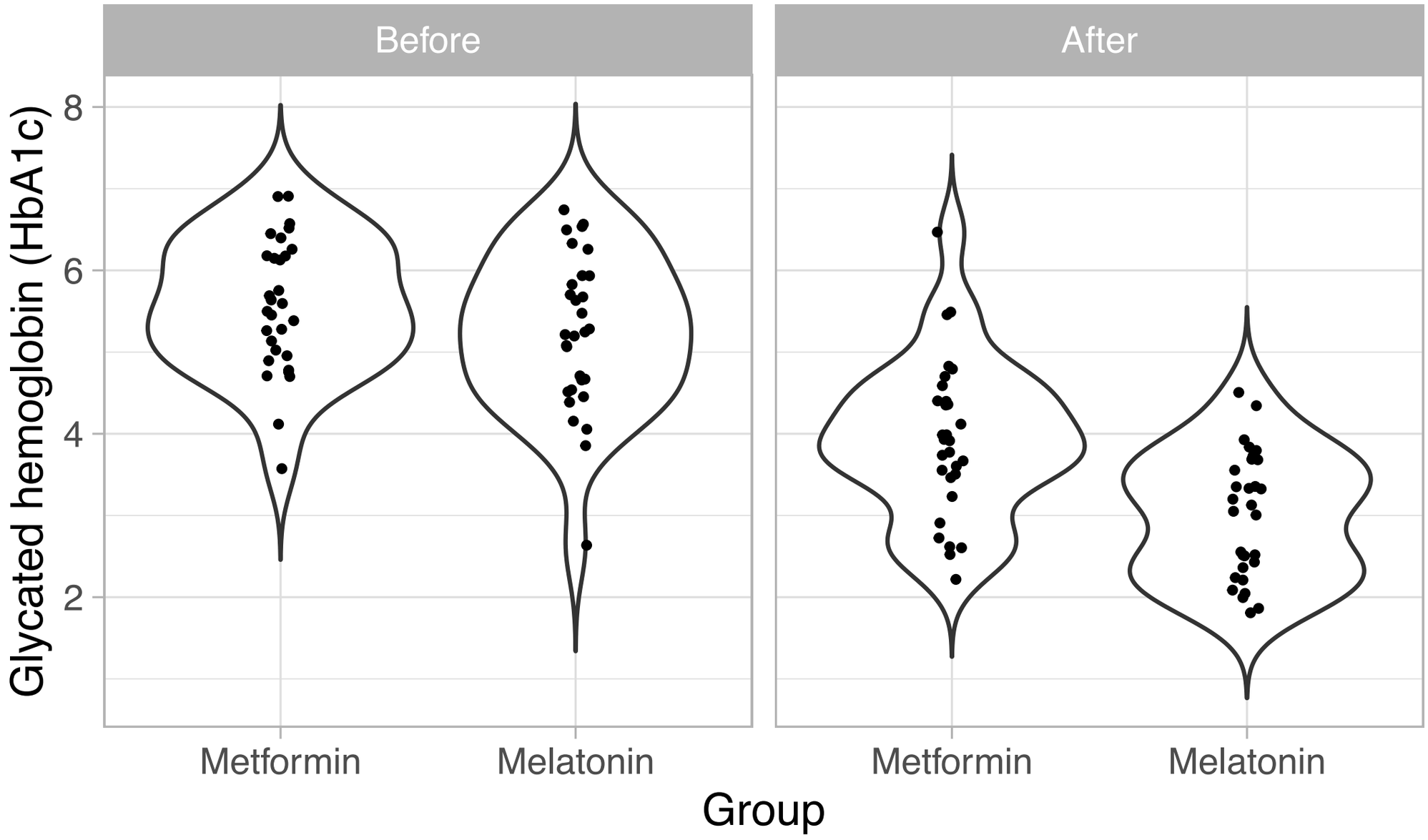

Input data
Define α
Normality test
Shapiro - Wilk
p < α
p > α
2 group
> 2 group
2 group
> 2 group
paired
unpaired
Kruskal - Wallis
t-test
p > α
Levene
p < α
Games - Howell
1 factor
2 factor
Wilcoxon
U Mann - Witney
Anova
2-way Anova
Significant?
p > α
p < α
Nothing significant
Post - hoc*
*Only for > 2 group tests
Power
Communicate
End
β > 0.2
β < 0.2

```
> shapiro.test(df$Proliferation)

    Shapiro-Wilk normality test

data:  df$Proliferation
W = 0.98079, p-value = 0.4631

> leveneTest(df$Proliferation~df$Group, center = mean)
Levenes Test Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  3  0.2246 0.8789
      56
```

```
> fit1 <- aov(df$Proliferation~df$Group)
> summary(fit1)
             Df Sum Sq Mean Sq F value Pr(>F)
df$Group      3   7.42   2.473   2.304 0.0867 .
Residuals    56  60.11   1.073
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(fit1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = df$Proliferation ~ df$Group)

$`df$Group`
                             diff         lwr      upr     p adj
Mel 0.1nM-Control  0.22916536 -0.77252460 1.230855 0.9298237
Mel 1nM-Control    0.28250133 -0.71918862 1.284191 0.8776541
Mel 10nM-Control   0.94481093 -0.05687902 1.946501 0.0712444
Mel 1nM-Mel 0.1nM  0.05333597 -0.94835398 1.055026 0.9989879
Mel 10nM-Mel 0.1nM 0.71564558 -0.28604438 1.717336 0.2432335
Mel 10nM-Mel 1nM   0.66230960 -0.33938035 1.664000 0.3077987
```


Proliferation marker
7.5
5.0
2.5
Control
Mel 0.1nM
Mel 1nM
Mel 10nM
Group

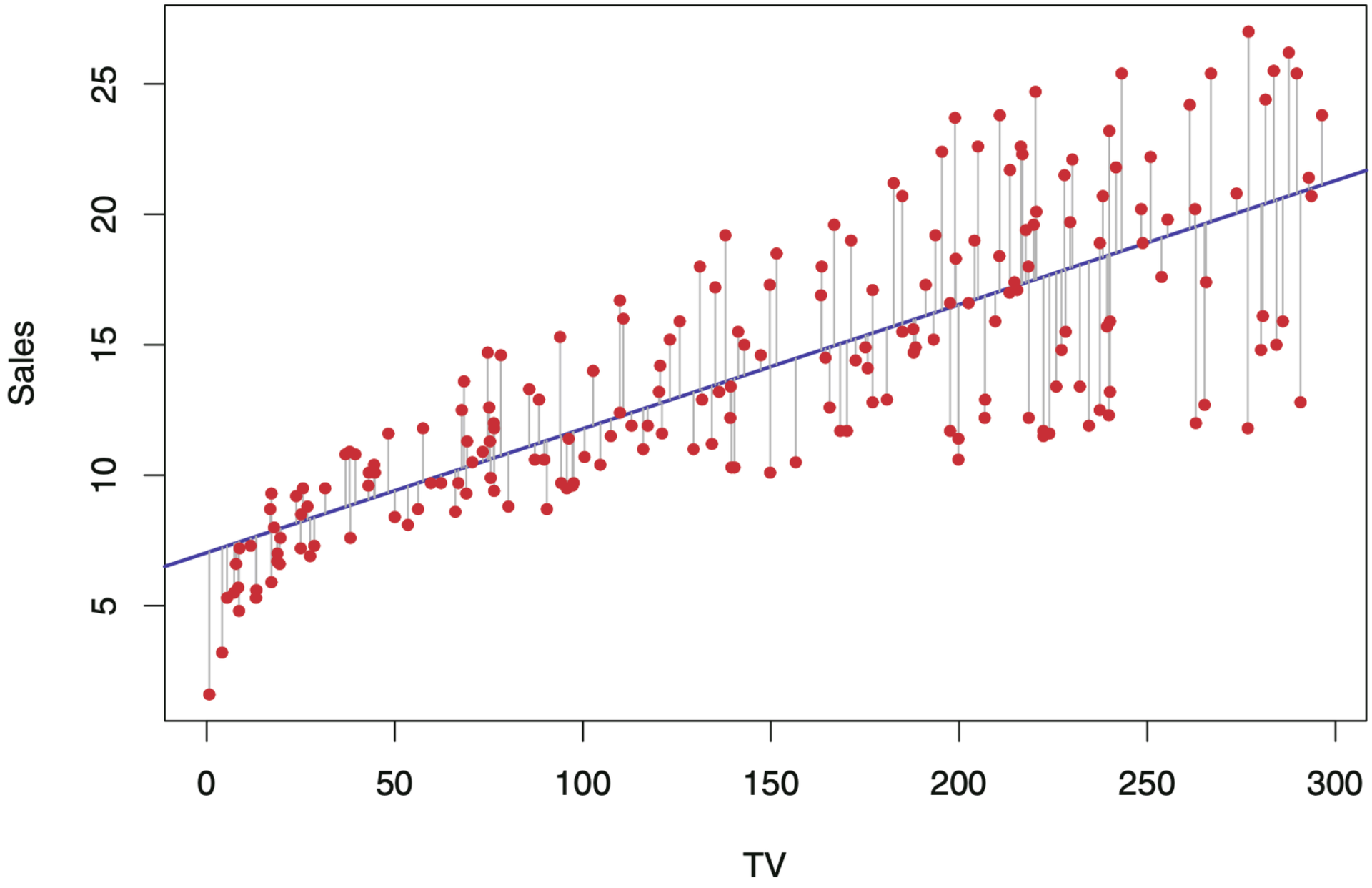

Input data
Define α
Normality test
Shapiro - Wilk
p < α
p > α
2 group
> 2 group
2 group
> 2 group
paired
unpaired
Kruskal - Wallis
t-test
p > α
Levene
p < α
Games - Howell
1 factor
2 factor
Wilcoxon
U Mann - Witney
Anova
2-way Anova
Significant?
p > α
p < α
Nothing significant
Post - hoc*
*Only for > 2 group tests
Power
Communicate
End
β > 0.2
β < 0.2


Before
After
Glycated hemoglobin (HbA1c)
8
6
4
2
Metformin
Melatonin
Metformin
Melatonin
Group

Regresión lineal simple

$$Y = b + mX$$

$$Y \approx \beta 0 + \beta 1 X$$



Regresión lineal simple

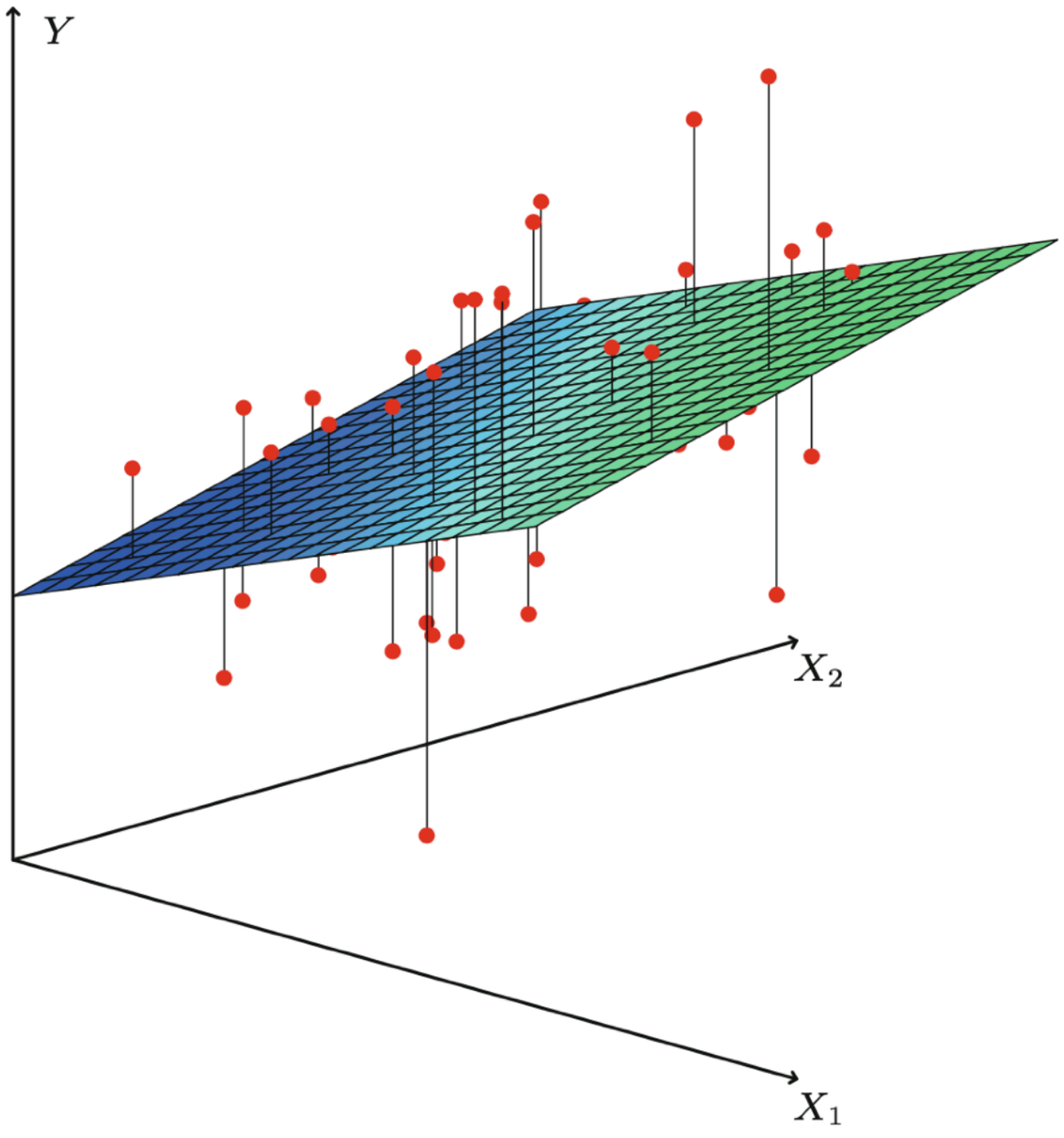
$$\text{Sales} \approx 7.0325 + 0.0475X$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

¿Cómo evaluamos la exactitud (*accuracy*) del modelo?

R^2 Statistic: Qué porcentaje de la varianza es explicada con los datos del modelo.

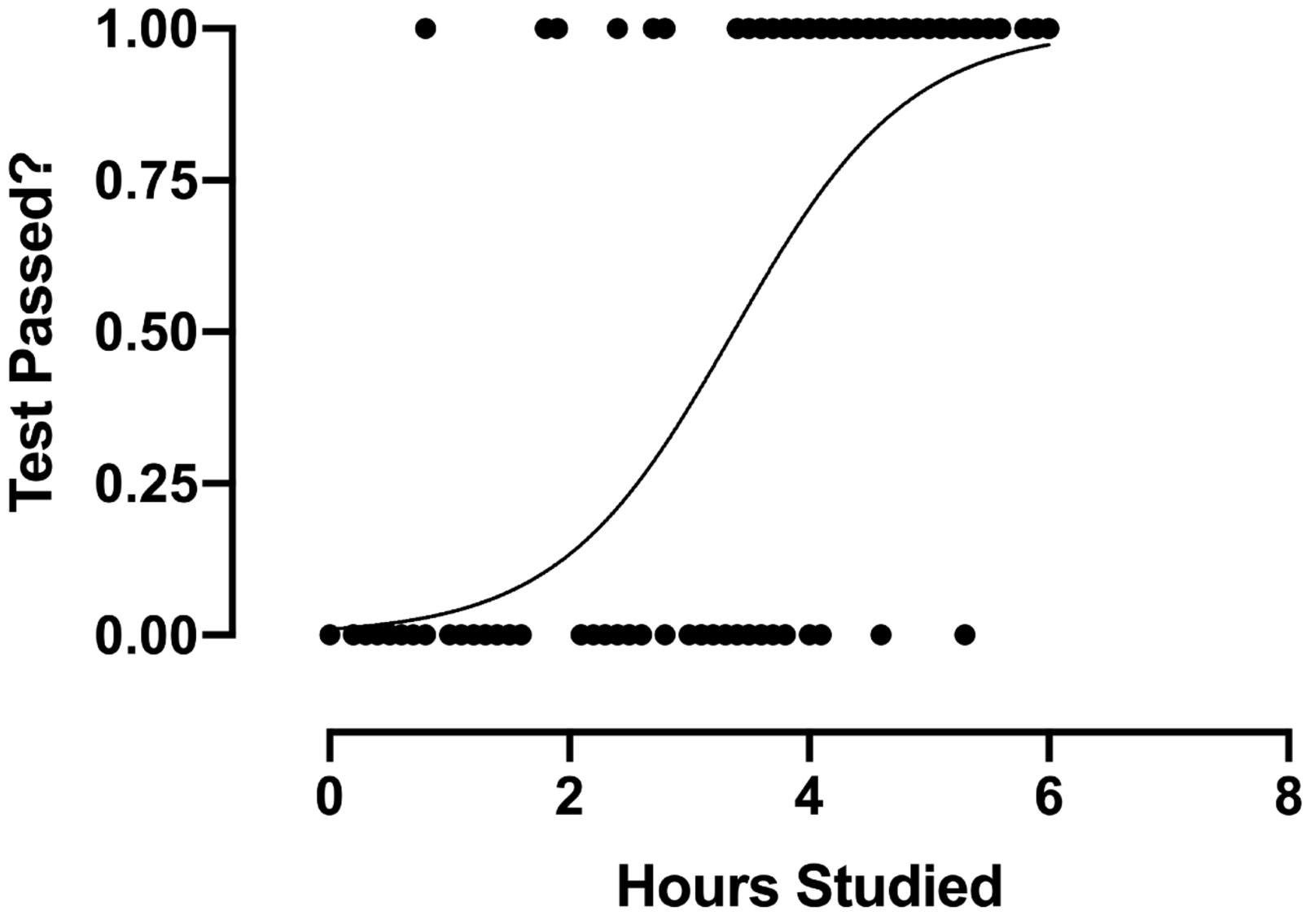
Regresión múltiple

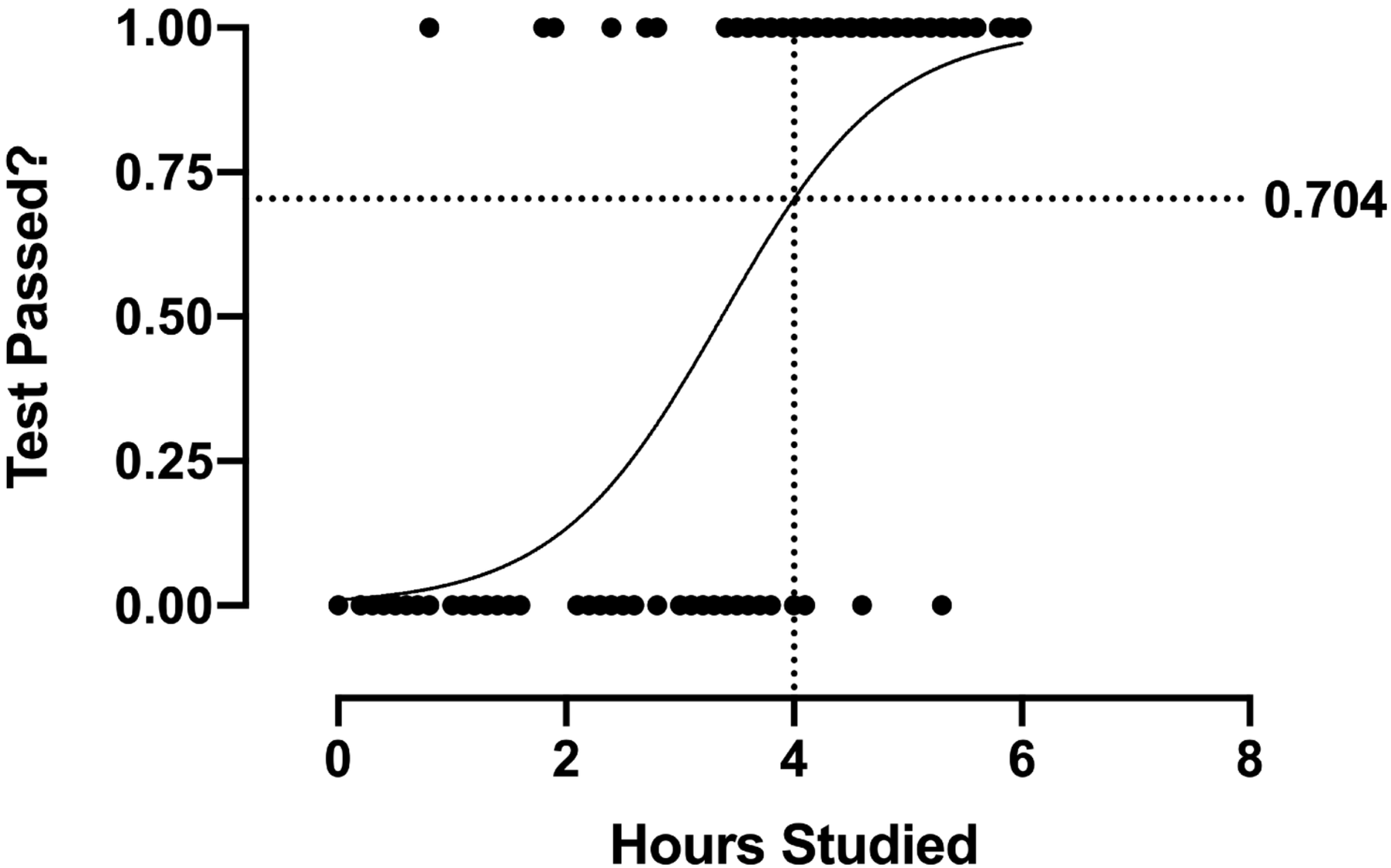


$$Y \approx \beta 0 + \beta 1X + \beta 2X + \ldots$$

Regresión logística

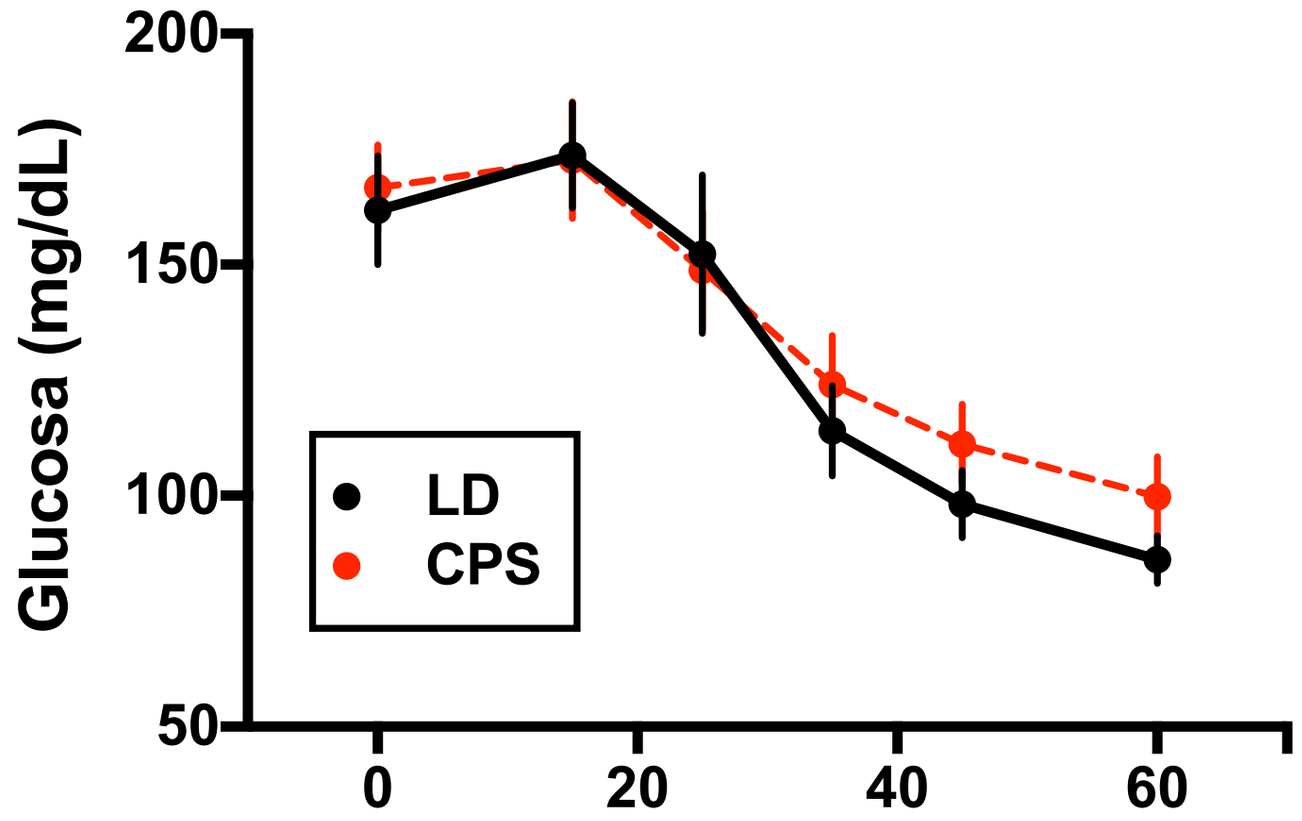
Respuesta binaria





$$OR = \frac{0.7 / 0.3}{0.3 / 0.7} = 8$$

Hay diferencias, pero no son significativas ¿existe algún error?



Potencia: 45%

¿Qué error podría estar cometiendo?