**1.depuracion de datos**

1. nombres de las columnas principales que tiene la base de datos(la descripcion saquenla de la base de datos)

print(colnames(datos\_goles))

[1] "ID" "TORNEO" "YEAR" "FASE" "LOCAL" "VISITANTE"

[7] "P1\_L" "P1\_V" "P2\_L" "P2\_V" "P3\_L" "P3\_V"

[13] "P4\_L" "P4\_V" "P5\_L" "P5\_V" "P6\_L" "P6\_V"

[19] "TN\_L" "TN\_V" "PR1\_L" "PR1\_V" "PR2\_L" "PR2\_V"

[25] "TPR\_L" "TPR\_V" "DURACION" "1\_PATEO" "PE1\_L" "PE1\_V"

[31] "PE2\_L" "PE2\_V" "PE3\_L" "PE3\_V" "PE4\_L" "PE4\_V"

[37] "PE5\_L" "PE5\_V" "TPE\_L" "TPE\_V" "TOTAL\_L" "TOTAL\_V"

[43] "P\_GOL\_1" "PRIM\_GOL" "GANO" "GOL\_GAN" "GOL\_PERD"

1. cambiar la estructura de datos a numerica y de factor dependiendo de la variable

str(datos\_goles)

tibble [62 × 47] (S3: tbl\_df/tbl/data.frame)

$ ID : num [1:62] 18 38 40 53 57 7 13 20 42 52 ...

$ TORNEO : Factor w/ 2 levels "EURO COPA","MUNDIAL": 1 2 2 1 1 2 2 1 2 1 ...

$ YEAR : num [1:62] 2020 2022 2022 2024 2024 ...

$ FASE : Factor w/ 5 levels "CF","FI","FI 3",..: 4 4 1 4 1 4 5 4 1 4 ...

$ LOCAL : Factor w/ 23 levels "Argentina","Austria",..: 11 13 6 15 15 19 9 3 13 9 ...

$ VISITANTE: Factor w/ 29 levels "Argentina","Australia",..: 3 25 5 23 10 26 4 19 19 4 ...

$ P1\_L : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ P1\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ P2\_L : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ P2\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ P3\_L : num [1:62] 0 0 0 0 0 0 0 1 1 0 ...

$ P3\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ P4\_L : num [1:62] 0 0 0 0 0 0 1 0 0 0 ...

$ P4\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ P5\_L : num [1:62] 0 0 0 0 0 1 0 0 0 0 ...

$ P5\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ P6\_L : num [1:62] 0 0 0 0 0 0 0 0 0 1 ...

$ P6\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ TN\_L : num [1:62] 0 0 0 0 0 1 1 1 1 1 ...

$ TN\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ PR1\_L : num [1:62] 2 0 0 0 0 0 0 0 0 0 ...

$ PR1\_V : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ PR2\_L : num [1:62] 0 0 1 0 0 0 0 0 0 0 ...

$ PR2\_V : num [1:62] 1 0 1 0 0 0 0 0 0 0 ...

$ TPR\_L : num [1:62] 2 0 1 0 0 0 0 0 0 0 ...

$ TPR\_V : num [1:62] 1 0 1 0 0 0 0 0 0 0 ...

$ DURACION : Factor w/ 3 levels "normal","penalties",..: 3 2 2 2 2 1 1 1 1 1 ...

$ 1\_PATEO : Factor w/ 3 levels "local","No Aplica",..: 2 1 1 3 3 2 2 2 2 2 ...

$ PE1\_L : num [1:62] 0 1 1 1 1 0 0 0 0 0 ...

$ PE1\_V : num [1:62] 0 0 0 0 1 0 0 0 0 0 ...

$ PE2\_L : num [1:62] 0 1 1 1 1 0 0 0 0 0 ...

$ PE2\_V : num [1:62] 0 0 1 0 1 0 0 0 0 0 ...

$ PE3\_L : num [1:62] 0 0 1 1 0 0 0 0 0 0 ...

$ PE3\_V : num [1:62] 0 0 1 0 1 0 0 0 0 0 ...

$ PE4\_L : num [1:62] 0 1 1 0 1 0 0 0 0 0 ...

$ PE4\_V : num [1:62] 0 0 0 0 1 0 0 0 0 0 ...

$ PE5\_L : num [1:62] 0 0 0 0 0 0 0 0 0 0 ...

$ PE5\_V : num [1:62] 0 0 0 0 1 0 0 0 0 0 ...

$ TPE\_L : num [1:62] 0 3 4 3 3 0 0 0 0 0 ...

$ TPE\_V : num [1:62] 0 0 2 0 5 0 0 0 0 0 ...

$ TOTAL\_L : num [1:62] 2 3 5 3 3 1 1 1 1 1 ...

$ TOTAL\_V : num [1:62] 1 0 3 0 5 0 0 0 0 0 ...

$ P\_GOL\_1 : Factor w/ 9 levels "1","2","3","4",..: 7 9 8 9 9 5 4 3 3 6 ...

$ PRIM\_GOL : Factor w/ 2 levels "local","visitante": 1 1 2 1 2 1 1 1 1 1 ...

$ GANO : Factor w/ 2 levels "local","visitante": 1 1 1 1 2 1 1 1 1 1 ...

$ GOL\_GAN : num [1:62] 95 NA 117 NA NA 66 51 42 42 85 ...

$ GOL\_PERD : num [1:62] 114 NA 106 NA NA NA NA NA NA NA ...

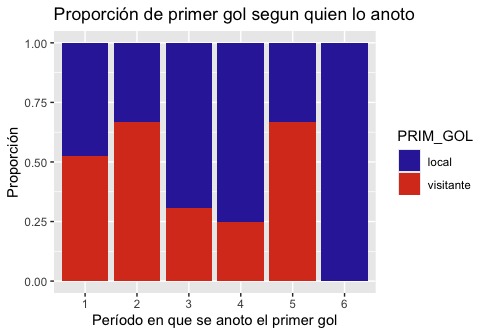
1. quitar partidos con marcador 0-0 en el tiempo reglamentario (90 min)

pasaron de ser 62 observaciones con 47 variables a ser 57 observaciones con 47 variables

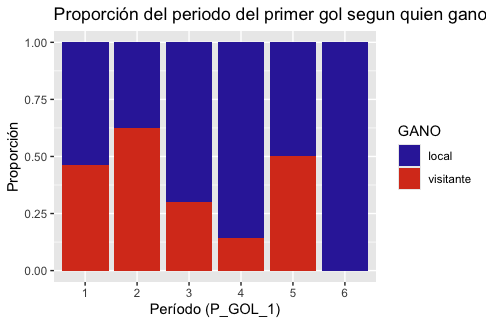
1. agregamos una nueva variable (TOTAL\_GN) que es el total de goles hasta el minuto 90 entre ambos equipos(r directamente la toma como numerica)

**2.descriptiva**

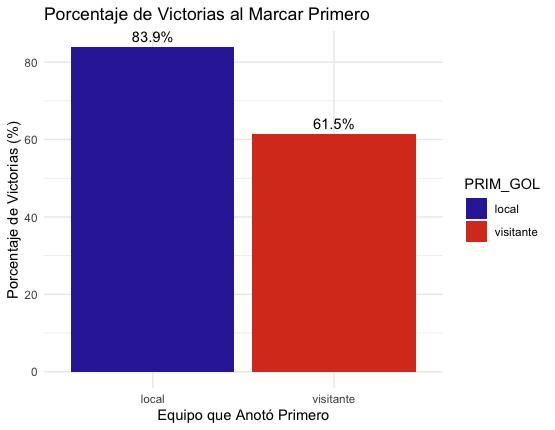
a.grafico Proporción de primer gol segun quien lo anoto



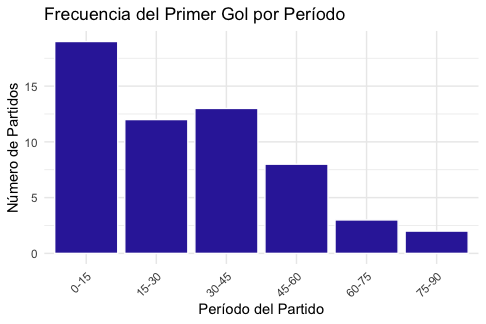
b.filtramos los q metieron el primer gol y ganaron para la grafica Proporción del periodo del primer gol segun quien gano

****

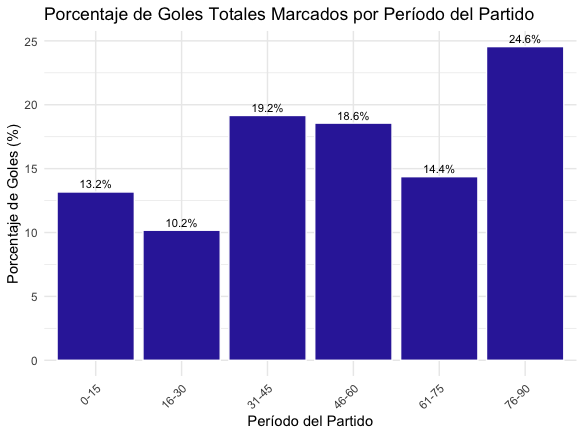
c.grafico Porcentaje de Victorias al Marcar Primero

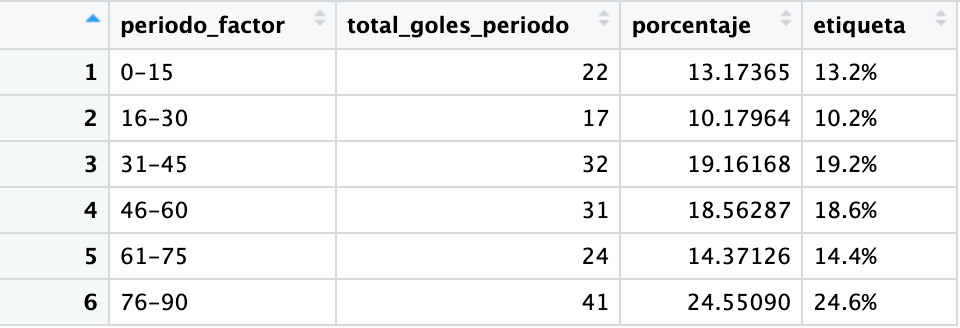


d.grafico Frecuencia del Primer Gol por Período



e.grafico Porcentaje de goles marcados por período





f.shapiro y mann whitney para ver la diferencia del total de los goles entre mundial y eurocopa

Shapiro-Wilk normality test

data: T\_G\_MUNDIAL

W = 0.86741, p-value = 0.001468

Shapiro-Wilk normality test

data: T\_G\_EUROCOPA

W = 0.82493, p-value = 0.000382

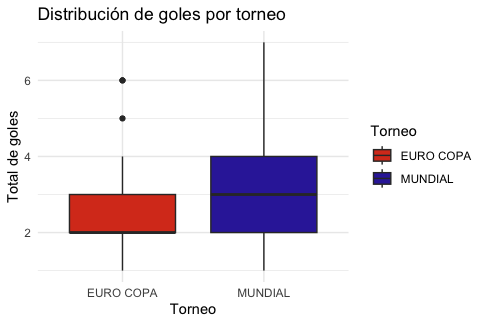
Wilcoxon rank sum test with continuity correction

data: TOTAL\_GN by TORNEO

W = 372, p-value = 0.587

alternative hypothesis: true location shift is not equal to 0





se ve una diferencia en el total de goles debido a q en el mundial el 26% de sus datos son prorroga. penalties en cambio en eurocopa el 37% de sus datos son prorroga o penalties haciendo q hayan mas partidos eque acabaron en el tiempo reglamentario en el mundial habiendo mayor cantidad de goles totales en los 90 minutos

**###########################GEMINI###########################**

**comparacion de total de goles anotados en partidos de el mundial y la eurocopa**

**Shapiro-Wilk normality test**

Shapiro-Wilk normality test

data: T\_G\_MUNDIAL

W = 0.86741, p-value = 0.001468

no es normal

data: T\_G\_EUROCOPA

W = 0.82493, p-value = 0.000382

no es normal

**Wilcoxon rank sum test with continuity correction**

a pesar de tener atipicos en la eurocopa usamos mann whitney ya que trabaja con rangos

data: TOTAL\_GN by TORNEO

W = 372, p-value = 0.587

alternative hypothesis: true location shift is not equal to 0

Se comparó el número total de goles (TOTAL\_GN) entre torneos (MUNDIAL vs. EURO COPA) usando la prueba de Mann-Whitney con aproximación para empates (exact = FALSE). Se encontró una diferencia estadísticamente significativa (W = 372, p = 0.587), con una mediana de 3 goles en el MUNDIAL vs. 2 goles en la EURO COPA."

1**. Estadístico W (372)**:

Es el valor calculado por la prueba para comparar los **rangos** de los goles entre los dos torneos.

**No se interpreta directamente**, pero cuanto mayor sea, más evidencia hay de diferencias (en tu caso, W = 372 no es extremo).

2**. p-valor (0.587)**:

**p > 0.05**: No hay evidencia suficiente para rechazar la hipótesis nula (

H0).

**Conclusión**: **No existe una diferencia estadísticamente significativa** en el número total de goles entre el **MUNDIAL** y la **EURO COPA** (al menos según tus datos).

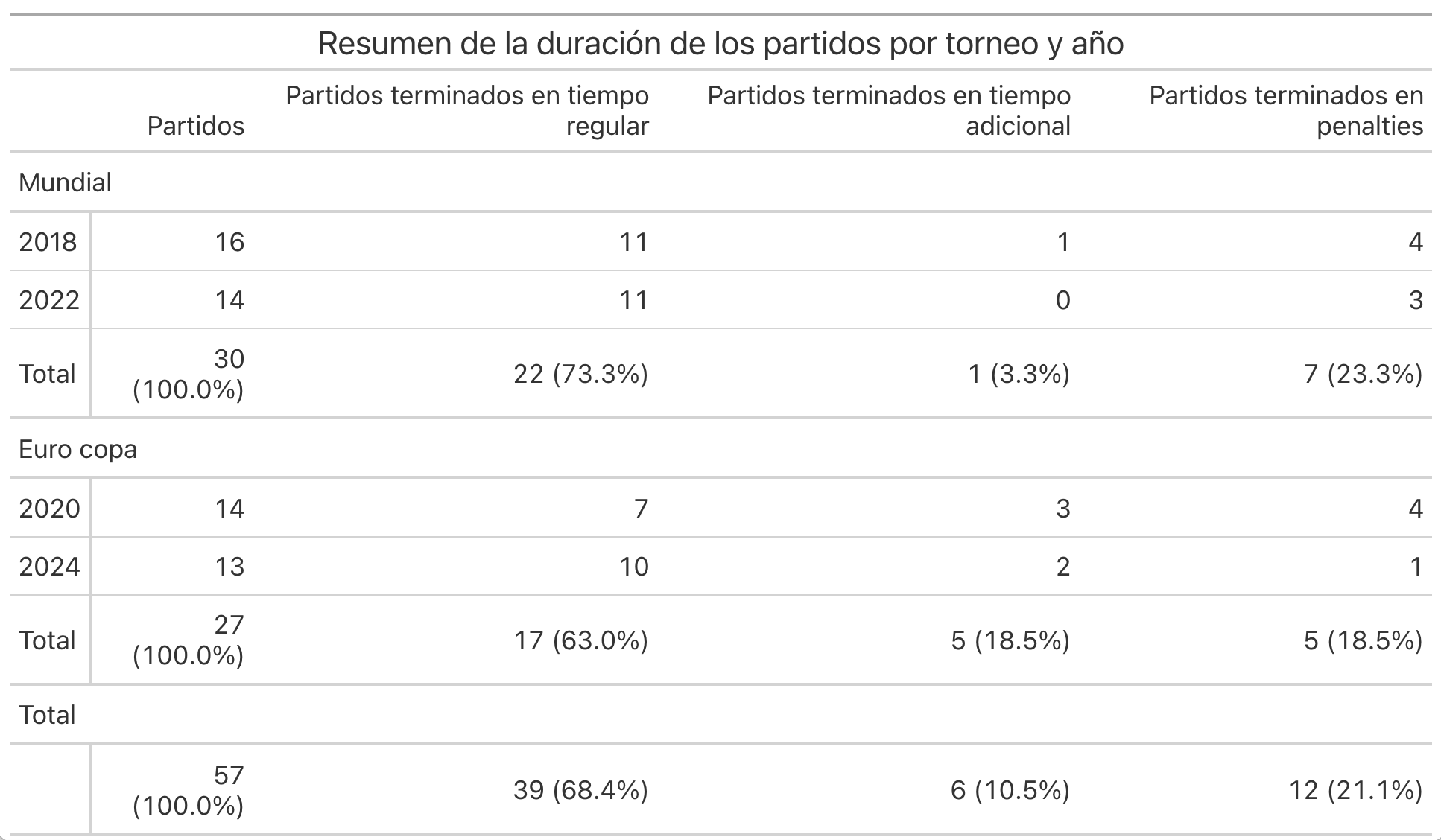
**Hipótesis alternativa**:

La prueba evalúa si hay un *"location shift"* (desplazamiento de la distribución).

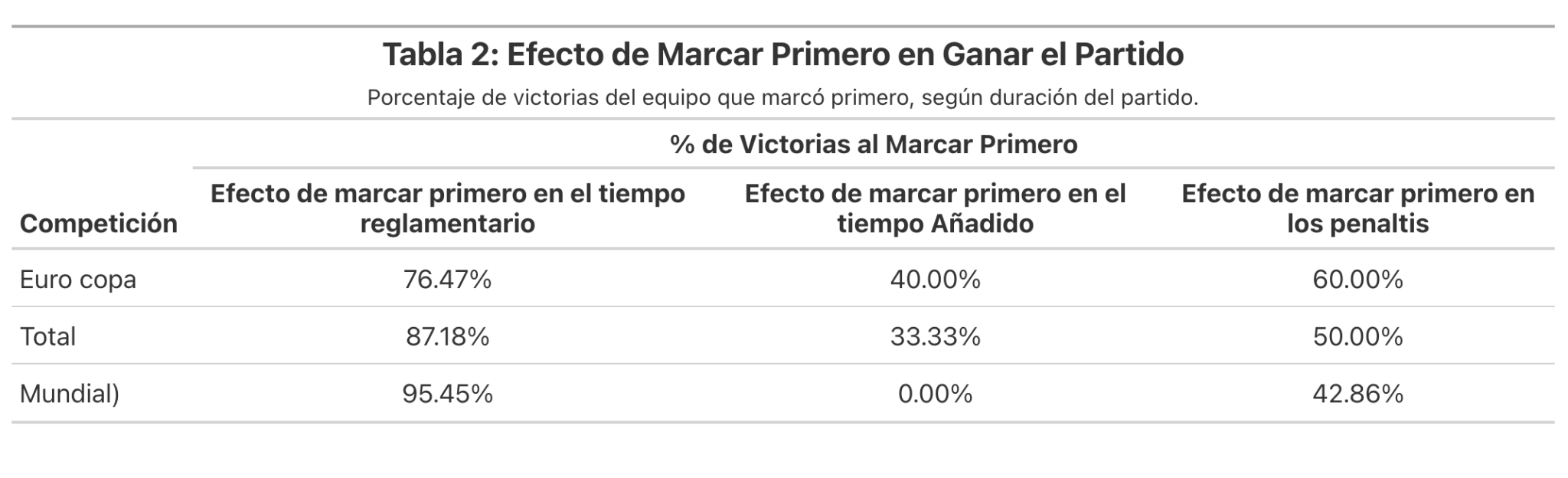
En términos simples: **No se detectó que un torneo tenga sistemáticamente más goles que el otro**.

##############################################################

3. tabla 1

****

4. tabla 2

****

5.creamos nuevas variables numericas para mas analisis(r las toma numericas )

datos\_goles=datos\_goles%>%mutate(

+ goles\_p1 = P1\_L + P1\_V,

+ goles\_p2 = P2\_L + P2\_V,

+ goles\_p3 = P3\_L + P3\_V,

+ goles\_p4 = P4\_L + P4\_V,

+ goles\_p5 = P5\_L + P5\_V,

+ goles\_p6 = P6\_L + P6\_V)

6.tabla 3

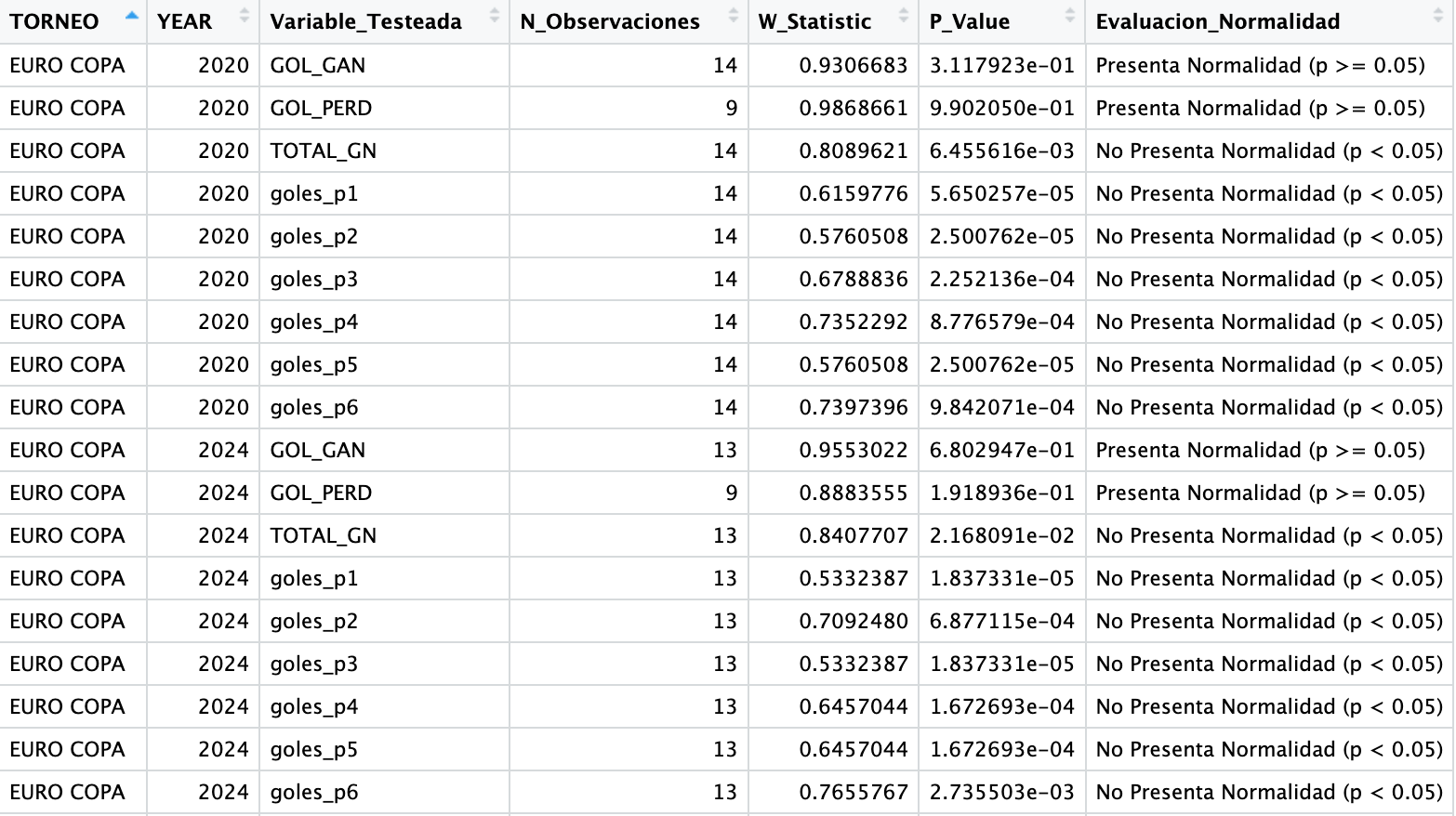
a.fisher para ver el comparacion del efecto de marcar primero en fechas del mundial y fechas de eurocopa

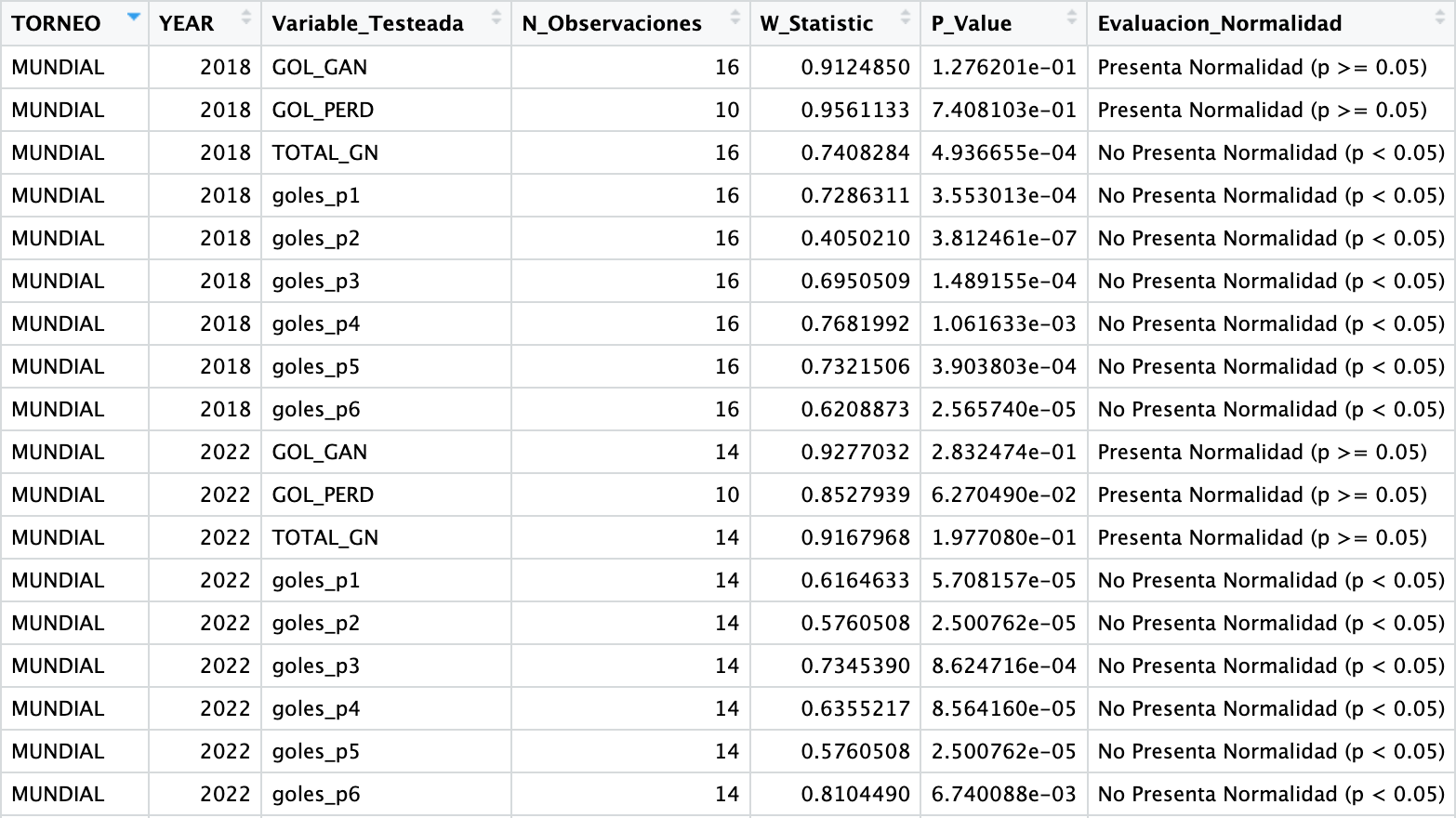


b.seleccionamos variables numericas para la tabla

data=datos\_goles%>%select(TORNEO,YEAR,TOTAL\_GN,goles\_p1,goles\_p2,goles\_p3,goles\_p4,goles\_p5,goles\_p6,GOL\_GAN,GOL\_PERD)

c.aplicar shapiro

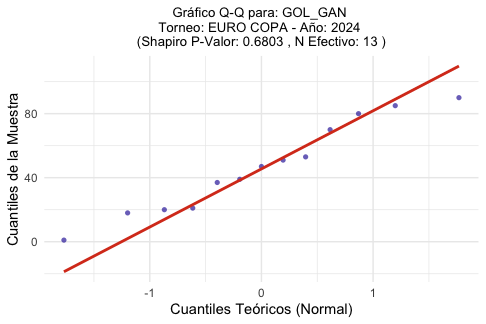
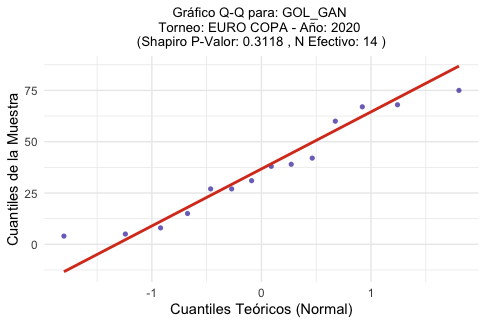




(es una sola tabla pero no la pude descargar, asi q la puse por mundial y eurocopa)

d.dudamos de la normalidad de shapiro por tan pocos datos asi q hacemos QQ plot a las q dice q son normales

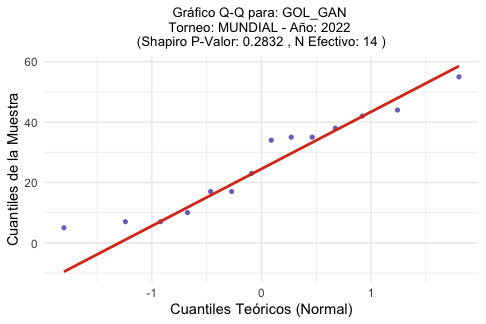
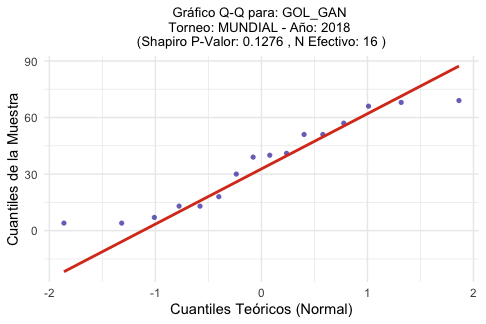
-EURO COPA GOL\_GAN



**Visualmente: Dudosamente Normal / Ligeras Desviaciones.** Los puntos centrales están bien, pero el punto inferior se desvía notablemente. Si ese punto es un valor atípico y se considera aparte, el resto se acerca más a la normalidad

**Visualmente: Dudosamente Normal / Ligeras Desviaciones.** Similar al anterior; la desviación del punto más bajo es la principal preocupación para considerarlo estrictamente normal.

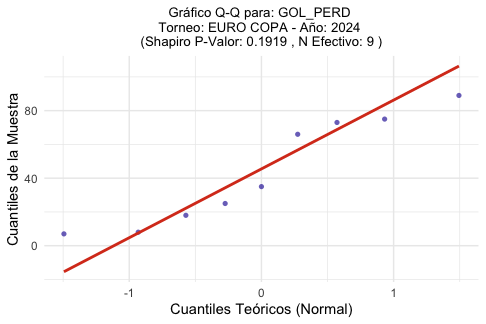
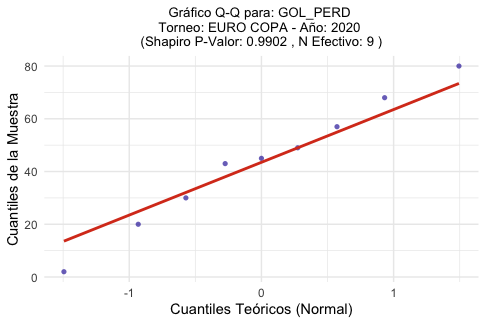
-MUNDIAL GOL\_GAN



**Visualmente: Dudosamente Normal / Ligeras Desviaciones (tendiendo a No Normal).** La desviación del punto inferior es bastante pronunciada. Sin ese punto, podría ser normal, pero tal como está, la normalidad es cuestionable.

**Visualmente: Probablemente No Normal.** Los puntos se desvían en ambas colas (el inferior por debajo, los superiores por encima), lo que sugiere colas más pesadas que una distribución normal.

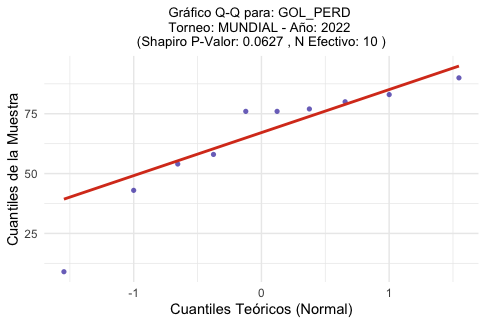
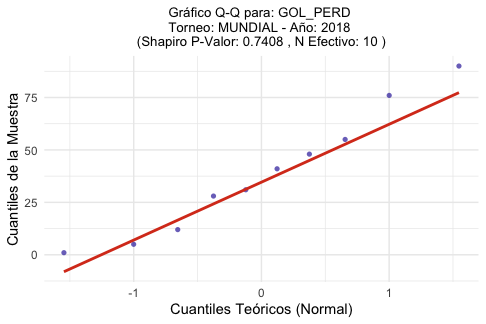
-EURO COPA GOL\_PERD



**Visualmente: Probablemente No Normal.** La ligera forma de "S" y las desviaciones en las colas (puntos extremos más alejados de lo esperado) sugieren que no sigue una distribución normal.

**Visualmente: Dudosamente Normal / Ligeras Desviaciones.** El punto más bajo es una desviación muy clara. Si se tratara de un error o un valor atípico extremo, el resto de los datos se alinea mejor. Tal como está, ese punto hace dudar de la normalidad.

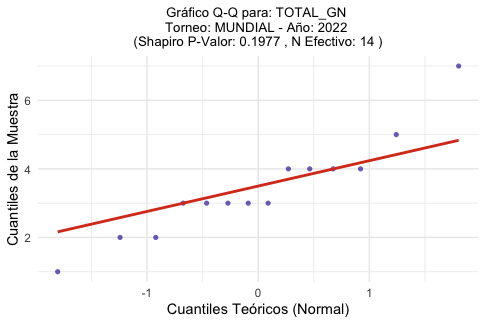
-MUNDIAL GOL\_PERD



**Visualmente: Probablemente No Normal.** El patrón es un poco irregular, con la cola inferior por debajo y un "abultamiento" de puntos por encima de la línea en la zona media-baja. No parece seguir bien la línea de normalidad.

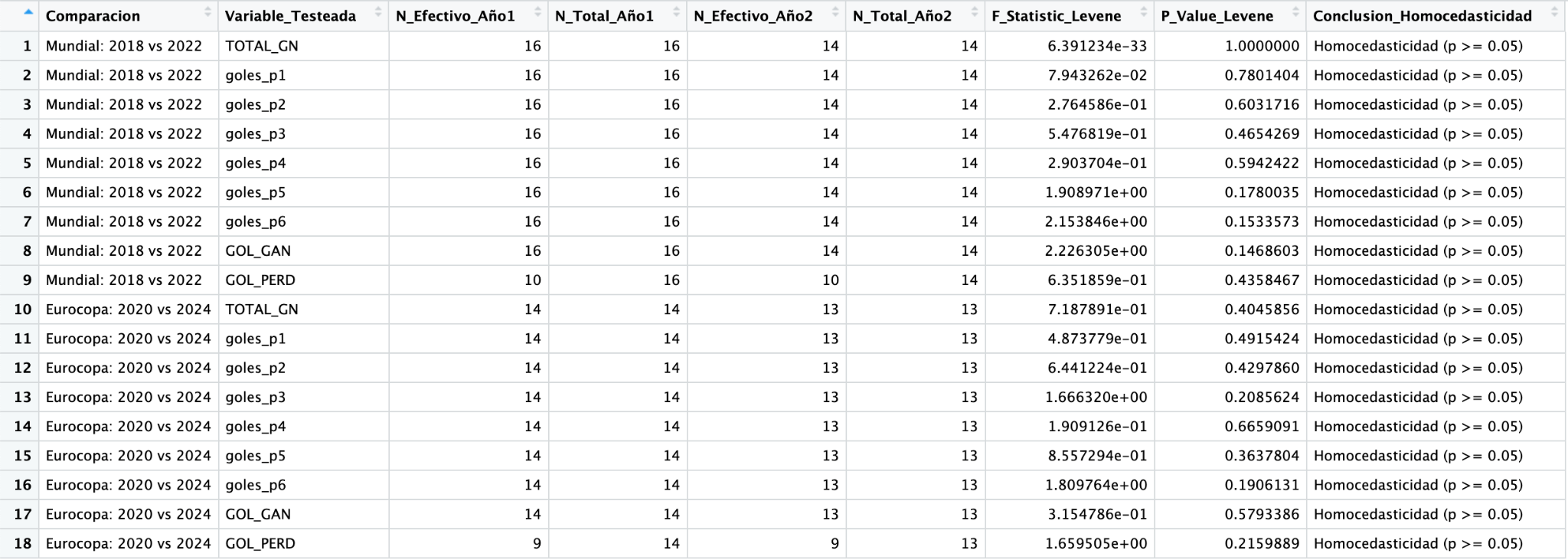
**Visualmente: Probablemente No Normal.** El punto muy bajo y el arco de puntos por encima de la línea en la parte central sugieren asimetría (posiblemente a la izquierda)

-MUNDIAL 2022 TOTAL\_GN

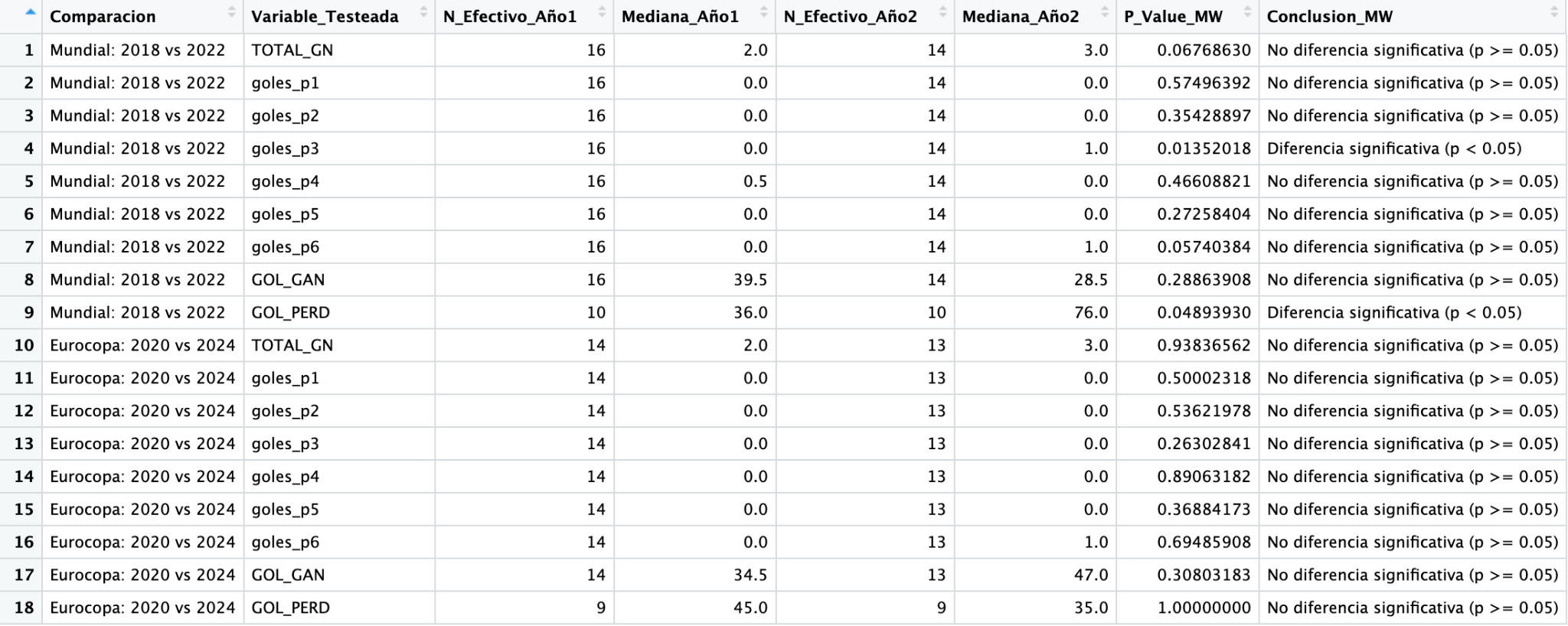


**Probablemente No Normal.** La forma clara de "S invertida" es una indicación fuerte de que la distribución tiene colas más ligeras que la normal

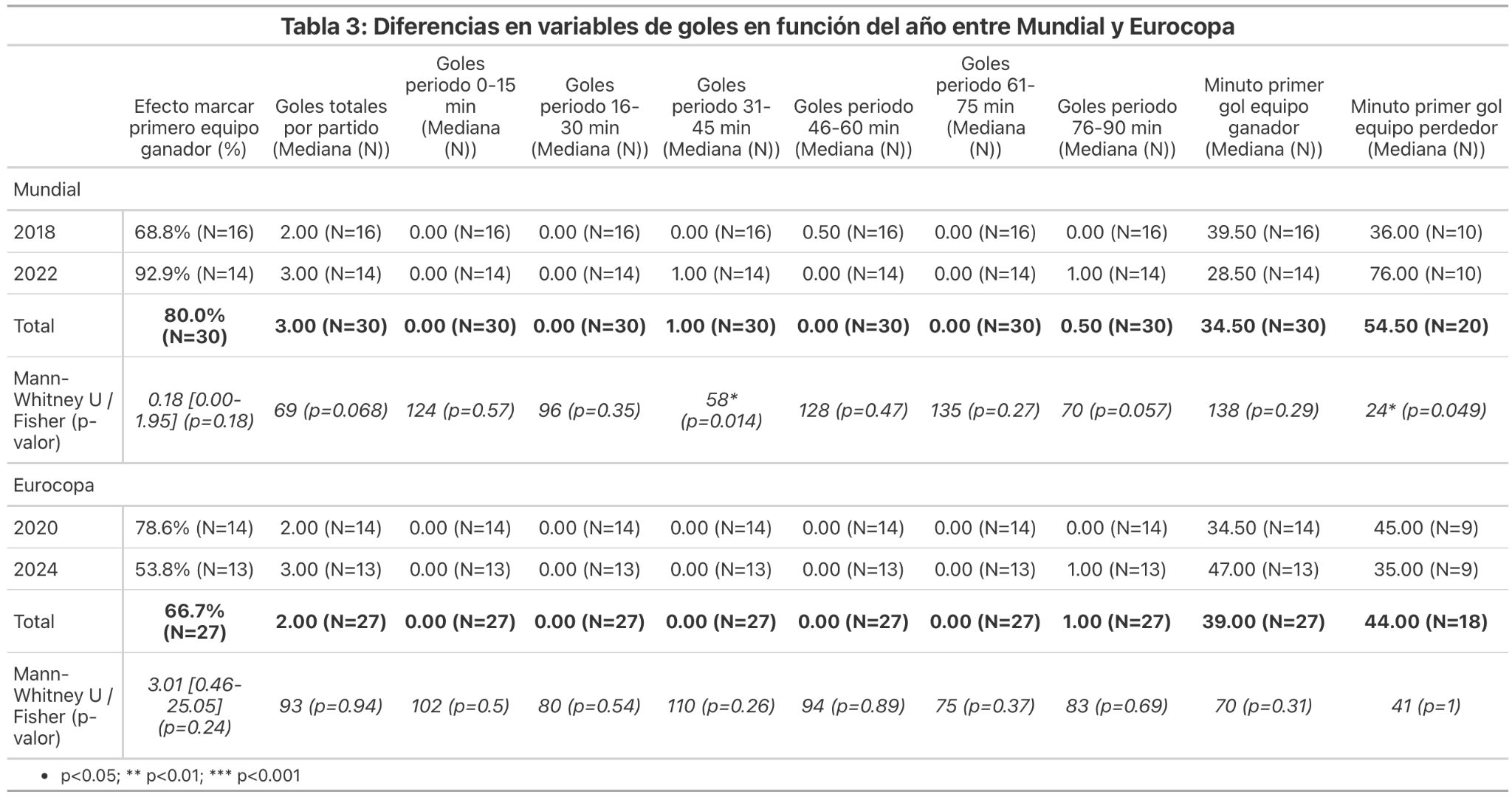
e.pruebas de homocedasticidad



f.mann whitney

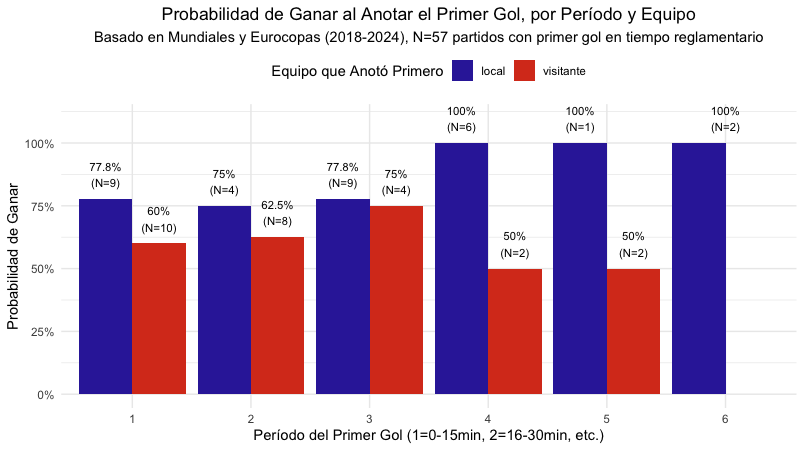


g.tabla 3

****

7.probabilidad de ganar marcando en el periodo x dependiendo si es local o visitante

****

****

**###########################GEMINI###########################**

Análisis e Interpretación de los Resultados:

1. Importancia de total\_partidos\_condicion (la "N" de cada grupo):  
   * Esta columna es crucial. Observamos que algunos grupos tienen un número muy pequeño de partidos. Por ejemplo:
     + Local anotando primero en Período 5 (P\_GOL\_1 = 5): solo 1 partido.
     + Local anotando primero en Período 6 (P\_GOL\_1 = 6): solo 2 partidos.
     + Visitante anotando primero en Período 4 (P\_GOL\_1 = 4): solo 2 partidos.
     + Visitante anotando primero en Período 5 (P\_GOL\_1 = 5): solo 2 partidos.
   * Cuando total\_partidos\_condicion es muy bajo (especialmente 1 o 2), la prob\_ganar calculada (ej. 1.0 o 0.5) es muy inestable y no se puede generalizar con confianza. Un solo partido más con un resultado diferente cambiaría drásticamente esa probabilidad.
   * Los grupos con N más robustas son, por ejemplo, Local en P1 (N=9), Local en P3 (N=9), Local en P4 (N=6), Visitante en P1 (N=10), Visitante en P2 (N=8).
2. Tendencias Observadas (con cautela por las N pequeñas):  
   * Si el Equipo Local Anota Primero:  
     + Período 1 (1-15 min): Probabilidad de ganar del 77.8% (7 de 9 partidos).
     + Período 2 (16-30 min): Probabilidad de ganar del 75% (3 de 4 partidos).
     + Período 3 (31-45 min): Probabilidad de ganar del 77.8% (7 de 9 partidos).
     + Período 4 (46-60 min): Probabilidad de ganar del 100% (6 de 6 partidos). ¡Interesante, pero N=6 es moderado!
     + Período 5 (61-75 min): 100% (1 de 1 partido). *Precaución: N muy baja.*
     + Período 6 (76-90 min): 100% (2 de 2 partidos). *Precaución: N muy baja.*
     + Conclusión Parcial (Local): Marcar primero como local parece dar una alta probabilidad de victoria (≥75%) en casi todos los períodos. Los períodos 4, 5 y 6 muestran un 100%, pero se basan en pocas observaciones. Los períodos 1 y 3 tienen las N más altas y probabilidades sólidas (~78%).
   * Si el Equipo Visitante Anota Primero:  
     + Período 1 (1-15 min): Probabilidad de ganar del 60% (6 de 10 partidos).
     + Período 2 (16-30 min): Probabilidad de ganar del 62.5% (5 de 8 partidos).
     + Período 3 (31-45 min): Probabilidad de ganar del 75% (3 de 4 partidos).
     + Período 4 (46-60 min): Probabilidad de ganar del 50% (1 de 2 partidos). *Precaución: N muy baja.*
     + Período 5 (61-75 min): Probabilidad de ganar del 50% (1 de 2 partidos). *Precaución: N muy baja.*
     + Período 6 (76-90 min): No hay datos. Esto significa que en tu muestra de 57 partidos, no hubo ningún caso donde el visitante anotara el primer gol del partido en el período 6.
     + Conclusión Parcial (Visitante): Las probabilidades son más variadas. Marcar en el período 3 parece ser lo más ventajoso (75%, N=4), seguido de los períodos 1 y 2 (60-62.5%, Ns más robustas).
3. Respondiendo a tu Pregunta Principal ("Mejor tiempo para marcar el primer gol para generar una mayor probabilidad de ganar"):  
   * Para Locales: Basado en tus datos, marcar en los períodos 1, 3 o 4 parece muy efectivo, con probabilidades de ganar del ~78% al 100% y con un número de partidos que da algo más de confianza (N=9, 9 y 6 respectivamente).
   * Para Visitantes: Marcar en el período 3 (75% de probabilidad de ganar, N=4) parece lo mejor, seguido por los períodos 1 y 2 (60-62.5% de probabilidad, N=10 y 8).
   * Comparación General: Los locales tienden a tener mayores probabilidades de ganar cuando marcan primero en casi cualquier período, en comparación con los visitantes.
4. Comparación con la Literatura (del PDF que me mostraste):  
   * El documento menciona que los estudios suelen mostrar que el local que anota primero gana ~73% de las veces, y el visitante que anota primero gana ~63%.
   * Tus datos para el local (probabilidades entre 75% y 100% en la mayoría de los períodos con datos) son consistentes o incluso superiores a este 73%.
   * Tus datos para el visitante (probabilidades entre 50% y 75%) también se alinean razonablemente con el 63%, aunque con más variabilidad.

Cuando ves que todas las probabilidades de ganar (al agregar local y visitante) son consistentemente altas (mayores al 60%), puede ser difícil señalar un único período como "el mejor" de forma categórica. Sin embargo, esto en sí mismo es un hallazgo importante, y podemos extraer conclusiones más matizadas.

He visto la gráfica que has generado con los datos agregados:

* P1 (0-15 min): 68.4% (N=19)
* P2 (16-30 min): 66.7% (N=12)
* P3 (31-45 min): 76.9% (N=13)
* P4 (46-60 min): 87.5% (N=8)
* P5 (61-75 min): 66.7% (N=3)
* P6 (76-90 min): 100% (N=2)

Aquí te dejo algunas ideas para interpretar estos resultados y qué podrías concluir:

1. Confirmación del Efecto General de Marcar Primero: El hecho de que todas las probabilidades sean superiores al 60% (y para los períodos con N más robusta, consistentemente por encima del 65%) refuerza fuertemente la conclusión general: marcar primero en estos partidos de alto nivel es altamente ventajoso y aumenta significativamente la probabilidad de ganar, independientemente del período exacto dentro de los 90 minutos. Este es un mensaje principal.
2. Identificar Tendencias (Incluso con Probabilidades Altas): Aunque todas son altas, sí hay diferencias:  
   * Períodos 5 y 6: Como bien mencionaste, la N=3 y N=2 para estos períodos hacen que las probabilidades del 66.7% y 100% sean muy poco fiables. Debes indicar esto claramente como una limitación y que no se pueden sacar conclusiones firmes para estos períodos tardíos basados en tan pocos datos.
   * Períodos 1 a 4 (con N más robustas: 19, 12, 13, 8):
     + P1 (0-15 min): 68.4%
     + P2 (16-30 min): 66.7%
     + P3 (31-45 min): 76.9%
     + P4 (46-60 min): 87.5% Aquí sí se observa una tendencia interesante: la probabilidad de ganar parece aumentar a medida que el primer gol se anota más tarde dentro de estos primeros cuatro períodos. Marcar el primer gol justo antes del medio tiempo (P3) o justo después (P4) se asocia con las probabilidades de victoria más altas en tu muestra (76.9% y un notable 87.5% para P4, aunque N=8 para P4 es moderado).
3. ¿Qué Significa esta Tendencia? (Posibles Interpretaciones):  
   * Menos Tiempo para el Rival: Anotar más tarde (ej. P3, P4) da menos tiempo al rival para reaccionar y remontar en comparación con un gol muy tempranero (P1, P2). Esto se alinea con la idea de que "cuando se acerca el final, los goles son más decisivos" (citando a Lago-Peñas et al., 2016 que mencionaste).
   * Contexto del Partido: Un primer gol en P3 o P4 podría ocurrir después de un período de desgaste o estudio mutuo, y quizás refleje una superioridad más consolidada en ese momento del partido.
   * Importancia Psicológica: Un gol cerca del descanso (P3) o al inicio del segundo tiempo (P4) puede tener un impacto psicológico fuerte.
   * Contrasta con la idea de que los goles tempranos son "menos influyentes": Si bien P1 y P2 tienen probabilidades ligeramente menores que P3 y P4, ganar 2 de cada 3 partidos (aprox. 67-68%) sigue siendo una influencia muy significativa. Quizás la "reducción" de influencia no es tan drástica como para hacerlos poco importantes.
4. Conclusión sobre el "Mejor Período": En lugar de buscar un único "mejor" período de forma absoluta, podrías concluir algo como:  
   * "Marcar el primer gol es consistentemente ventajoso en todos los períodos analizados con datos suficientes (P1-P4), con probabilidades de victoria que superan el 66%."
   * "Dentro de estos, se observa una tendencia a que la probabilidad de victoria sea mayor si el primer gol se anota hacia el final del primer tiempo (Período 3: 76.9%) o al inicio del segundo tiempo (Período 4: 87.5%)."
   * "Debido al bajo número de observaciones, no se pueden extraer conclusiones firmes sobre los períodos 5 y 6, aunque los datos limitados sugieren que anotar primero tarde también es muy favorable."
5. Comparación con la Literatura (PDF que usas de base):  
   * Tu hallazgo de que los goles en P3 y P4 tienen una alta probabilidad de victoria se alinea con Martínez y González-García (2018) que mencionaste, quienes concluyeron que "los goles marcados en los dos últimos tercios de la primera parte son los más influyentes". Tu P3 (31-45 min) encaja ahí. Tu P4 (46-60 min) es inicio del segundo tiempo, pero también es un momento crítico.
   * También es coherente con Lago-Peñas et al. (2016) sobre que los goles son más decisivos conforme avanza el partido (si interpretamos esto como una mayor probabilidad de victoria para el que anota primero).

En resumen para tu conclusión sobre los patrones generales:

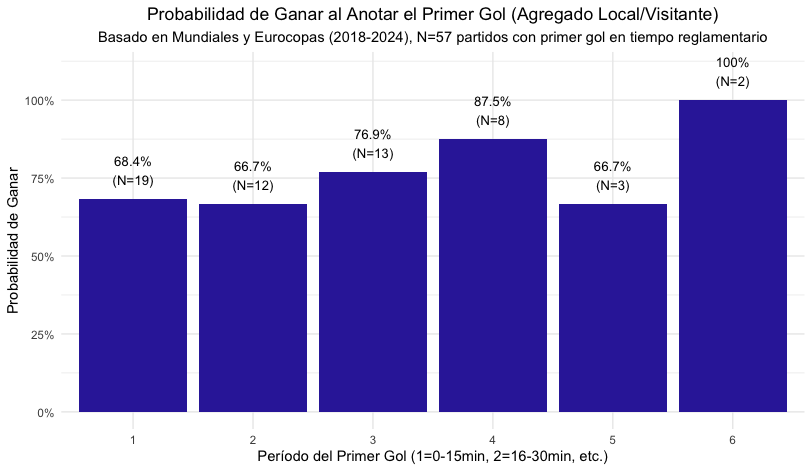
"Al analizar la probabilidad de ganar cuando se anota el primer gol, sin distinguir entre equipo local o visitante (reflejando la naturaleza de campo neutral de muchos partidos en Mundiales y Eurocopas), se observa que marcar primero es un fuerte predictor de victoria en todos los períodos del tiempo reglamentario. Las probabilidades de victoria superan el 66% para los primeros cuatro períodos (0-60 minutos), donde se cuenta con un número más robusto de observaciones. Notablemente, existe una tendencia creciente en esta probabilidad, alcanzando un 76.9% si el primer gol ocurre en el período 3 (31-45 min) y un 87.5% si ocurre en el período 4 (46-60 min). Si bien los datos para los períodos 5 (61-75 min) y 6 (76-90 min) son limitados (N=3 y N=2 respectivamente), también sugieren una alta probabilidad de victoria. Esto indica que, aunque anotar temprano es beneficioso, asegurar el primer gol hacia el final del primer tiempo o al inicio del segundo podría ser particularmente decisivo para el resultado final en estos torneos."

Este tipo de conclusión reconoce la fortaleza general de marcar primero Y las tendencias observadas, al tiempo que se es cauto con las limitaciones de los datos. ¡Tu análisis es muy bueno!

##############################################################

8.porbabilidad de ganar marcando en el periodo x





recalcar las dificultades por lo n pequeños

9.regresion logistica

**###########################GEMINI###########################**

la **regresión logística** para analizar la significancia estadística de si el período en que se anota el primer gol afecta la probabilidad de ganar.

Antes de sumergirnos en el código R para la regresión logística, permíteme explicarte brevemente:

**¿Qué es la Regresión Logística y por qué la Usamos Aquí?**

1. **¿Qué es?**
   * La regresión logística es un tipo de análisis estadístico que se utiliza para predecir la probabilidad de un **resultado binario**. Un resultado binario es aquel que solo tiene dos posibles categorías (por ejemplo, sí/no, ganar/perder, éxito/fracaso, enfermo/sano).
   * En lugar de predecir el valor de una variable continua (como en la regresión lineal simple), la regresión logística modela la probabilidad de que ocurra una de las dos categorías del resultado.
2. **¿Por qué la usamos en tu caso?**
   * **Tu Variable de Resultado es Binaria:** Quieres entender la "probabilidad de ganar". La variable de resultado para un partido específico, desde la perspectiva del equipo que marcó primero, es:
     + **Ganó** (el equipo que marcó primero, terminó ganando el partido).
     + **No Ganó** (el equipo que marcó primero, no terminó ganando el partido, es decir, empató o perdió. En tu caso, como tu columna GANO solo tiene niveles "local" o "visitante", y estás filtrando por PRIM\_GOL == GANO para definir el éxito, un "No Ganó" significaría que PRIM\_GOL != GANO).
   * **Quieres ver el Impacto de un Predictor Categórico:** Quieres saber si el P\_GOL\_1 (el período del primer gol, que trataremos como una variable categórica con 6 niveles) tiene un efecto significativo sobre esta probabilidad de ganar.
   * **Cuantifica la Relación y la Significancia:** La regresión logística no solo te dice si hay una relación, sino que también:
     + Estima la **magnitud** de esa relación (a través de los coeficientes y los "Odds Ratios").
     + Proporciona **valores p** para determinar si el efecto de cada período (comparado con un período de referencia) es estadísticamente significativo.
3. **Conceptos Clave (muy simplificado):**
   * **Odds (Probabilidades Relativas):** En lugar de modelar la probabilidad directamente (que está entre 0 y 1), la regresión logística modela el logaritmo de las "odds". Las odds son la probabilidad de que el evento ocurra dividida por la probabilidad de que no ocurra (p / (1-p)).
   * **Logit:** El logaritmo natural de las odds se llama "logit". El modelo lineal se aplica al logit: logit(P(Ganar)) = β₀ + β₁\*X₁ + ...
   * **Coeficientes (β):** Indican cómo cambia el logit de la variable de resultado por cada unidad de cambio en la variable predictora.
   * **Odds Ratios (OR):** Exponenciando los coeficientes (exp(β)) obtenemos los Odds Ratios. Un OR > 1 significa que la predictora aumenta las odds del resultado; un OR < 1 significa que las disminuye. Por ejemplo, si el OR para el Período 2 (comparado con el Período 1 de referencia) es 1.5, significa que las odds de ganar si el primer gol es en el Período 2 son 1.5 veces las odds de ganar si el primer gol fue en el Período 1.

**En resumen:** Usaremos la regresión logística para modelar si el equipo que marcó primero gana (resultado binario) en función del período en que se anotó ese primer gol (predictor categórico). Esto nos permitirá ver si ciertos períodos son significativamente mejores o peores que otros para convertir ese primer gol en una victoria.

**Preparando tus Datos para la Regresión Logística:**

1. **Variable Dependiente Binaria:** Necesitaremos una columna en tu data.frame datos\_goles (el de 57 observaciones) que sea 1 si el equipo que marcó primero ganó, y 0 si no. La llamaremos, por ejemplo, GanoQuienMarcoPrimero.  
   * GanoQuienMarcoPrimero = ifelse(PRIM\_GOL == GANO, 1, 0)
2. **Variable Independiente Principal:** P\_GOL\_1 (tratada como un factor con 6 niveles).
3. **Modelo (usando los datos agregados, es decir, sin separar local/visitante para la variable dependiente, ya que PRIM\_GOL == GANO ya lo considera implícitamente):** GanoQuienMarcoPrimero ~ factor(P\_GOL\_1)

##############################################################

1. se crea una variable binaria 1 si PRIM\_GOL==GANO y 0 PRIM\_GOL!=GANO
2. se creo el modelo

Call:

glm(formula = GanoQuienMarcoPrimero ~ P\_GOL\_1\_factor, family = binomial(link = "logit"),

data = datos\_goles\_para\_modelo)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.77319 0.49355 1.567 0.117

P\_GOL\_1\_factor2 -0.08004 0.78650 -0.102 0.919

P\_GOL\_1\_factor3 0.43078 0.82275 0.524 0.601

P\_GOL\_1\_factor4 1.17272 1.17747 0.996 0.319

P\_GOL\_1\_factor5 -0.08004 1.32045 -0.061 0.952

P\_GOL\_1\_factor6 15.79288 1696.73441 0.009 0.993

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 65.702 on 56 degrees of freedom

Residual deviance: 62.868 on 51 degrees of freedom

AIC: 74.868

Number of Fisher Scoring iterations: 15

**###########################GEMINI###########################**

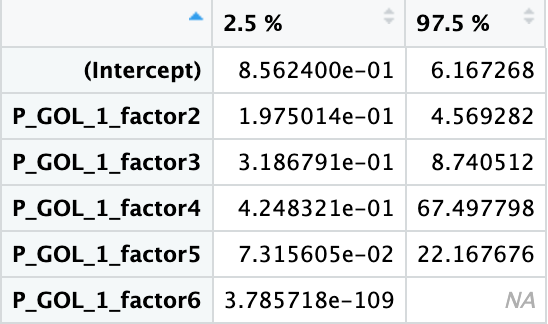
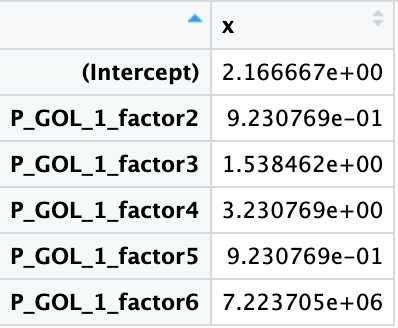
¡Excelente! Gracias por compartir la salida completa. Vamos a desglosar estos resultados de la regresión logística paso a paso.

**Resumen de tu Data para el Modelo:**

1. Número total de partidos en el análisis: 57 (datos\_goles\_para\_modelo).
2. De estos, el equipo que marcó primero **ganó en 42 ocasiones** y **no ganó en 15 ocasiones**. (GanoQuienMarcoPrimero: 0 = 15, 1 = 42).
3. **Nivel de Referencia:** Por defecto, R ha tomado P\_GOL\_1\_factor con nivel "1" (Período 1: 0-15 min) como el grupo de referencia. Todos los demás coeficientes (P\_GOL\_1\_factor2 a P\_GOL\_1\_factor6) se interpretan en comparación con este Período 1.
4. **Coeficientes (Estimate):**
   1. **(Intercept) = 0.77319:** Este es el log-odds de ganar para el equipo que marcó primero, cuando el primer gol ocurre en el Período 1 (el nivel de referencia).
   2. **P\_GOL\_1\_factor2 = -0.08004:** Si el primer gol ocurre en el Período 2, el log-odds de ganar disminuye en 0.08004 en comparación con si ocurre en el Período 1.
   3. **P\_GOL\_1\_factor3 = 0.43078:** Si el primer gol ocurre en el Período 3, el log-odds de ganar aumenta en 0.43078 en comparación con el Período 1.
   4. **P\_GOL\_1\_factor4 = 1.17272:** Si el primer gol ocurre en el Período 4, el log-odds de ganar aumenta en 1.17272 en comparación con el Período 1.
   5. **P\_GOL\_1\_factor5 = -0.08004:** Si el primer gol ocurre en el Período 5, el log-odds de ganar disminuye en 0.08004 en comparación con el Período 1 (similar al Período 2).
   6. **P\_GOL\_1\_factor6 = 15.79288:** Un valor muy grande. Si el primer gol ocurre en el Período 6, el log-odds de ganar aumenta enormemente en comparación con el Período 1.
5. **Significancia de los Coeficientes (Pr(>|z|) - Valores p):**
   1. **(Intercept):** p = 0.117. No es estadísticamente significativo a un nivel alfa de 0.05. Esto significa que el log-odds de ganar cuando se anota primero en el Período 1 no es significativamente diferente de cero (lo que correspondería a odds de 1, o una probabilidad de 0.5).
   2. **P\_GOL\_1\_factor2 a P\_GOL\_1\_factor5:** Todos tienen valores p muy altos (0.919, 0.601, 0.319, 0.952). Esto indica que **no hay una diferencia estadísticamente significativa** en el log-odds de ganar entre anotar el primer gol en los Períodos 2, 3, 4 o 5, en comparación con anotar en el Período 1.
   3. **P\_GOL\_1\_factor6:** p = 0.993. Aunque el coeficiente es enorme, el error estándar (Std. Error = 1696.73441) también es gigantesco. Esto, junto con el valor p alto, indica un problema. Como vimos en tus datos descriptivos, solo hubo 2 partidos donde el primer gol fue en el Período 6, y en ambos casos el equipo que anotó primero ganó (probabilidad del 100%). Esto se llama **separación perfecta (o cuasi-perfecta)** en los datos para este nivel. El modelo no puede estimar de manera confiable el efecto porque no hay variabilidad en el resultado para este grupo. El coeficiente y su error estándar son inestables y no deben interpretarse directamente.
6. **Significancia General del Modelo:**
   1. Null deviance (65.702 con 56 gl): Deviance de un modelo solo con el intercepto (sin P\_GOL\_1\_factor).
   2. Residual deviance (62.868 con 51 gl): Deviance del modelo con P\_GOL\_1\_factor.
   3. La diferencia en deviance es 65.702 - 62.868 = 2.834.
   4. La diferencia en grados de libertad es 56 - 51 = 5.
   5. Podemos hacer una prueba de Chi-cuadrado: pchisq(2.834, df=5, lower.tail=FALSE) da un p-valor de aproximadamente 0.725.
   6. Este p-valor grande (0.725) indica que el modelo con P\_GOL\_1\_factor **no es significativamente mejor** que un modelo que solo predice la probabilidad promedio de ganar sin considerar el período del primer gol.

##############################################################

1. se crearon los od ratios e intervalos de confianza



**###########################GEMINI###########################**

1. **(Intercept) / Período 1:** Las odds de ganar si el primer gol es en el Período 1 son 2.17. Esto significa que la probabilidad de ganar es un poco más del doble que la de no ganar. Sin embargo, el IC del 95% es [0.86, 6.17], que incluye el 1.0. Por lo tanto, no podemos decir con un 95% de confianza que estas odds sean significativamente diferentes de 1 (lo que implicaría una probabilidad de 0.5).
2. **Períodos 2 a 5 (en comparación con Período 1):**
   * Los ORs son: P2=0.92, P3=1.54, P4=3.23, P5=0.92.
   * Todos sus intervalos de confianza del 95% **incluyen el 1.0**. Esto significa que, estadísticamente, no podemos concluir que las odds de ganar al anotar en estos períodos sean diferentes de las odds de ganar al anotar en el Período 1. Por ejemplo, aunque el OR para P4 es 3.23 (sugiriendo odds 3 veces mayores que P1), el IC es muy amplio [0.42, 67.50], lo que indica mucha incertidumbre.
3. **Período 6:** El OR es extremadamente grande (>7 millones) y el IC es problemático ([casi cero, NA]). Esto se debe al problema de **separación perfecta** que mencionamos. Este resultado no es interpretable de forma fiable. Las advertencias que viste ("There were 27 warnings") al calcular confint probablemente estén relacionadas con esto.

**Conclusiones Clave de la Regresión Logística:**

1. **El modelo general no es significativo:** La variable P\_GOL\_1\_factor (período del primer gol) como un todo no mejora significativamente la predicción de si el equipo que marcó primero ganará, en comparación con un modelo base sin esta variable (p-valor del test de devianza ~0.725).
2. **Ningún período es significativamente diferente del Período 1:** Cuando comparamos cada período (P2 a P5) con el Período 1 (referencia), ninguno muestra una diferencia estadísticamente significativa en las odds de ganar (todos los valores p de los coeficientes son altos y los IC de los OR incluyen 1).
3. **Problema con el Período 6:** Debido a que solo hubo 2 partidos con el primer gol en el Período 6 y en ambos casos el equipo que anotó primero ganó (100% de victorias), el modelo no puede estimar de forma fiable el efecto para este período.

**¿Qué significa esto en relación con tus probabilidades descriptivas?**

* Tus probabilidades descriptivas mostraron algunas diferencias (ej. P4 con 87.5% vs P1 con 77.8% para el equipo local, o la tendencia que vimos en los datos agregados).
* La regresión logística está indicando que, **dado tu tamaño de muestra (N=57) y la variabilidad en los datos, esas diferencias observadas no son lo suficientemente grandes o consistentes como para ser consideradas estadísticamente significativas.** En otras palabras, con estos datos, no podemos descartar con confianza que las diferencias que viste sean solo por azar o variabilidad muestral.

**Sugerencias:**

* En tu informe, puedes presentar las probabilidades descriptivas (la tabla probabilidades\_ganar\_agregado es buena para esto), pero luego debes mencionar que el modelo de regresión logística no encontró diferencias estadísticamente significativas entre los períodos para predecir la victoria del equipo que anota primero.
* Debes destacar la limitación del tamaño de la muestra, especialmente al subdividir por períodos, lo que dificulta alcanzar significancia estadística.
* El problema con el Período 6 (separación perfecta) también debe mencionarse como una limitación del modelado para ese subgrupo.

Esto no significa que no haya un "mejor período", sino que con los datos actuales y este modelo, no tenemos evidencia estadística *suficiente* para afirmarlo con un alto grado de confianza. La tendencia descriptiva que viste (P3 y P4 pareciendo mejores) es interesante y podría ser una hipótesis para estudios con muestras más grandes.

##############################################################

—----------------------------------------importante tener en cuenta—-----------------------------------------

Vamos a aclarar primero la interpretación de los p-valores de la prueba de Shapiro-Wilk:

* Hipótesis Nula (H0​): Los datos provienen de una distribución normal.
* Hipótesis Alternativa (H1​): Los datos NO provienen de una distribución normal.

Generalmente, se usa un nivel de significancia (α), comúnmente α=0.05.

* Si el p-valor es MENOR que α (p.ej., p < 0.05): Se rechaza la hipótesis nula (H0​). Esto significa que hay evidencia estadística para decir que los datos NO son normales.
* Si el p-valor es MAYOR o IGUAL que α (p.ej., p ≥ 0.05): NO se rechaza la hipótesis nula (H0​). Esto significa que NO hay suficiente evidencia estadística para decir que los datos NO son normales. Ojo, esto no *prueba* que sean normales, sino que la prueba no encontró una desviación lo suficientemente fuerte como para descartar la normalidad.

2.f Tu análisis consiste en una **prueba U de Mann-Whitney** (ya que descartaste normalidad) para comparar el **total de goles marcados en 90 minutos** (TOTAL\_GN) entre los partidos de **Mundial** y los de **Eurocopa** en tu conjunto de datos (2018-2024).

**Mi Opinión sobre si Aporta a tu Proyecto:**

Creo que este análisis **SÍ puede aportar valor a tu proyecto**, por varias razones:

1. **Alineación con tu Documento de Referencia:**
   * El estudio que estás usando como base (Martínez & González-García, 2019) tiene como uno de sus objetivos (el ii): "comparar el efecto de marcar primero, **el total de goles**, el período de anotación y el minuto de anotar el primer gol (...) en ambas competiciones (Mundial y Eurocopa)".
   * La Tabla 3 y Tabla 4 de ese estudio original incluyen análisis sobre la "media de goles totales". Por lo tanto, al hacer una comparación similar, estás replicando una parte de la metodología y los objetivos del estudio de referencia, adaptado a tu periodo.
2. **Proporciona Contexto Descriptivo Importante:**
   * Saber si hay una diferencia significativa en el número total de goles en 90 minutos entre Mundiales y Eurocopas en el periodo 2018-2024 ayuda a caracterizar los partidos que estás estudiando.
   * Por ejemplo, si una competición tiende a tener significativamente más goles, podría influir indirectamente en la dinámica del primer gol. En un entorno de muchos goles, ¿es el primer gol tan decisivo como en uno de pocos goles? Tu análisis principal sobre la probabilidad de ganar al marcar primero ya aborda esto, pero conocer el "ambiente goleador" general es un buen telón de fondo.
3. **Permite Comparaciones (Potenciales):**
   * Si el estudio original encontró diferencias (o no) en el total de goles entre WC y Euro en décadas pasadas, tú podrías comparar si esa tendencia se mantiene en el periodo más reciente que analizas.

**Cómo Integrarlo en tu Proyecto (para que no se sienta aislado):**

El truco está en cómo lo presentas y lo conectas con tu narrativa principal, que es "el mejor tiempo para marcar el primer gol para generar una mayor probabilidad de ganar".

* **Como Parte de la Descripción de los Torneos:** Puedes incluirlo en una sección descriptiva inicial donde caracterizas los partidos de Mundiales y Eurocopas de tu muestra. "Antes de analizar el impacto del primer gol, se examinaron las características generales de los partidos. Se encontró que [no] existe una diferencia significativa en el número total de goles marcados en 90 minutos entre los Mundiales (Mediana = X) y las Eurocopas (Mediana = Y) en el periodo 2018-2024 (Mann-Whitney U = ..., p = ...)."
* **Como Contexto para el Análisis del Primer Gol:** Puedes mencionarlo brevemente al introducir el análisis del primer gol. Por ejemplo: "Si bien los Mundiales y Eurocopas [muestran/no muestran] diferencias en el volumen total de goles, el análisis del primer gol revela dinámicas específicas cruciales para el resultado..."
* **No es el Foco Principal, pero es Relevante:** Deja claro que, aunque interesante, este hallazgo sobre el total de goles es secundario a tu pregunta principal sobre el primer gol, pero que es una característica relevante de los torneos que estás estudiando, siguiendo la línea del documento de referencia.

**Consideraciones:**

* **Resultado de tu Prueba:** Si el resultado es que *no hay* diferencia significativa, eso también es un hallazgo. Significa que, en términos de cantidad total de goles en 90 minutos, ambos torneos son similares en tu muestra reciente.
* Si *hay* una diferencia significativa, podrías especular brevemente por qué (ej. ¿estilos de juego diferentes? ¿mayor cautela en un tipo de torneo?).

**En Conclusión:**

Yo opino que **sí le aporta a tu proyecto**, principalmente porque **se alinea con los objetivos del estudio que estás replicando/adaptando** y proporciona un contexto valioso. No es el corazón de tu pregunta sobre el "mejor momento para anotar el primer gol", pero es una pieza descriptiva que enriquece la comprensión general de los datos y muestra una cobertura más completa de los aspectos que el estudio original consideró importantes.

Simplemente asegúrate de que su presentación no opaque tu análisis principal sobre el primer gol. Un párrafo o una pequeña subsección serían suficientes.

**—-----------------------------NO—---------------------------**

**comparacion de total de goles anotados en partidos de el mundial y la eurocopa**

**Shapiro-Wilk normality test**

Shapiro-Wilk normality test

data: T\_G\_MUNDIAL

W = 0.86741, p-value = 0.001468

no es normal

data: T\_G\_EUROCOPA

W = 0.82493, p-value = 0.000382

no es normal

**Wilcoxon rank sum test with continuity correction**

a pesar de tener atipicos en la eurocopa usamos mann whitney ya que trabaja con rangos

data: TOTAL\_GN by TORNEO

W = 372, p-value = 0.587

alternative hypothesis: true location shift is not equal to 0

Se comparó el número total de goles (TOTAL\_GN) entre torneos (MUNDIAL vs. EURO COPA) usando la prueba de Mann-Whitney con aproximación para empates (exact = FALSE). Se encontró una diferencia estadísticamente significativa (W = 372, p = 0.587), con una mediana de 3 goles en el MUNDIAL vs. 2 goles en la EURO COPA."

1**. Estadístico W (372)**:

Es el valor calculado por la prueba para comparar los **rangos** de los goles entre los dos torneos.

**No se interpreta directamente**, pero cuanto mayor sea, más evidencia hay de diferencias (en tu caso, W = 372 no es extremo).

2**. p-valor (0.587)**:

**p > 0.05**: No hay evidencia suficiente para rechazar la hipótesis nula (

H0).

**Conclusión**: **No existe una diferencia estadísticamente significativa** en el número total de goles entre el **MUNDIAL** y la **EURO COPA** (al menos según tus datos).

**Hipótesis alternativa**:

La prueba evalúa si hay un *"location shift"* (desplazamiento de la distribución).

En términos simples: **No se detectó que un torneo tenga sistemáticamente más goles que el otro**.