# Parcial computacional sobre Test de Hipótesis

## Afinidad de Género

Borrell Trinidad, DNI: 43245969

*Métodos Estadísticos en Física Experimental*

2023

## 1. Distribuciones

En el primer inciso se nos pide realizar una simulación en la que se asigna $N = 10000$ veces unx candidatx ($m$ o $h$) con unx directorx ($M$ o $H$). Y con los resultados obtenidos calcular las distribuciones de $P(M|m)$ y $P(H|h)$ que describen la probabilidad de haber tenido unx directorx mujer u hombre, si se tiene unx candidatx mujer u hombre, respectivamente. Las consideraciones a tener en cuenta para realizar esta simulación fueron que, en primer lugar, se cumple la hipótesis nula $H_0$, la cual establece que no hay un sesgo de género por parte de lxs candidatxs a la hora de elegir directorx. Y en segundo lugar, se consideró que el análisis debía realizarse anualmente de forma independiente ya que, las condiciones de contorno variaban año a año y consecuentemente las probabilidades condicionales posibles.

La lógica de la simulación entonces fue generar dos listas que describían al conjunto de candidatxs y directorxs de ese año. Luego usando una función particular mezclarlas de forma aleatoria y hacer las asignaciones correspondientes al orden resultante de las listas. De allí contabilizar los éxitos, es decir, en este caso serían cuando hay una mujer como candidata y directora, y tambien cuando hay un hombre como candidato y director (obviamente serían dos tipos de éxitos). La aletoriedad del orden de las listas garantizaba que se cumpla $H_0$. Finalmente con los éxitos contabilizados se calculan las probabilidades condicionales usando que $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Los resultados obtenidos para las $N = 10000$ iteraciones permitieron obtener las distribuciones de $P(M|m)$ y $P(H|h)$, las cuales se pueden ver en las figuras 1) y 2) respectivamente.

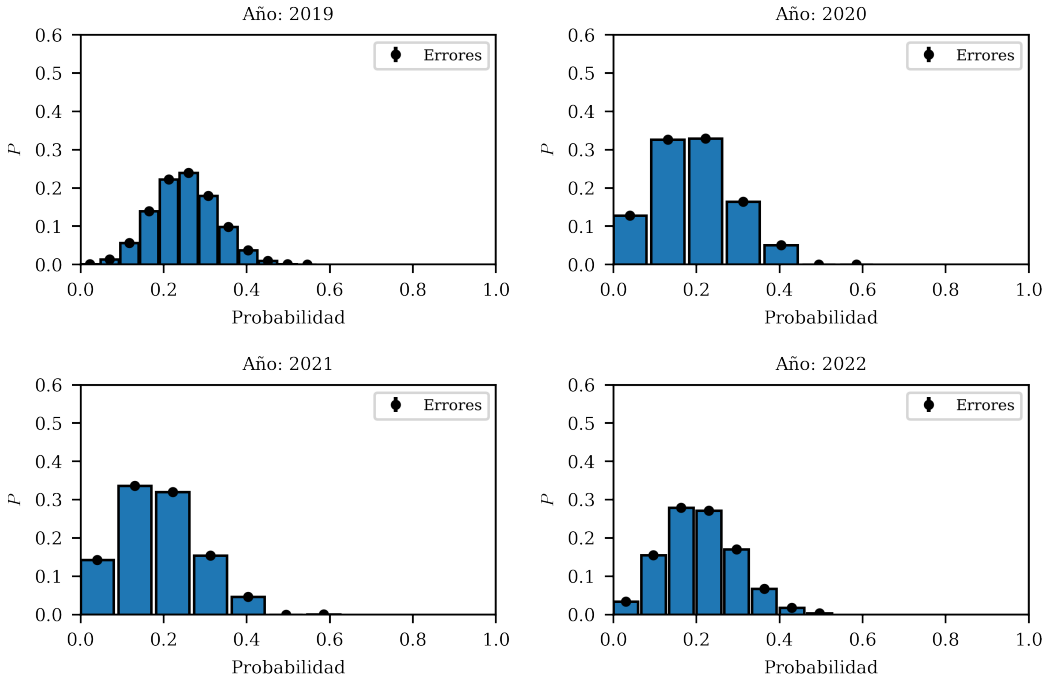



Figura 1: Distribución de la probabilidad condicional $P(M|m)$ año a año con sus correspondientes errores.

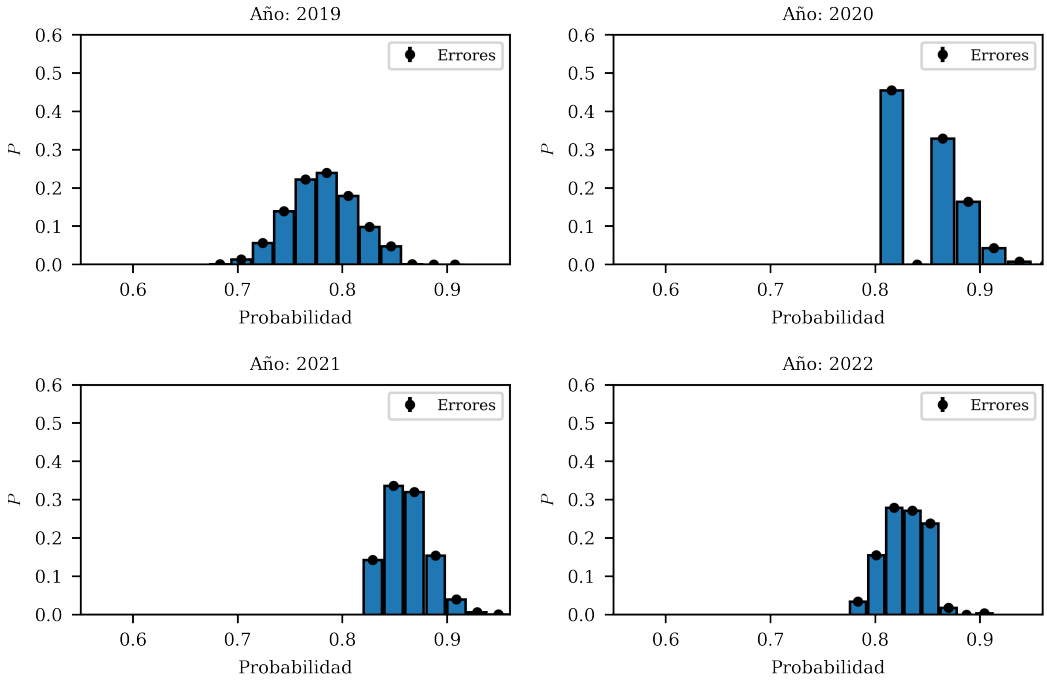



Figura 2: Distribución de la probabilidad condicional $P(H|h)$ año a año con sus correspondientes errores.

En primer lugar, para ambos casos los errores se establecieron considerando que cada bin contaba con una distribución poissoniana, cuya $\mu$ es la altura del bin y por lo tanto $\sigma = \sqrt{\mu}$. Por otro lado, se observa que las distribuciones de $P(M|m)$ tienen todas eventos en valores menores a 0,6, es decir, no se observó ningún caso en que la probabilidad condicional sea mayor a 0,6, mientras que para el caso de $P(H|h)$ estas se encuentran todos los resultados arriba de 0,7.

## 2. Definiendo el test

Para encontrar las probabilidades condicionales críticas de cada año con una significancia $\alpha = 0{,}05$, lo que se hizo fue considerar que la región de rechazo se encontraba en la zona derecha de la distribución ya que si se obtienen probabilidades condicionales muy altas esto podría deberse a una cierta afinidad por el propio género (cosa que justamente queremos verificar). Luego, para calcular la probabilidad crítica con la significancia correspondiente, la idea sería sumar los bines desde la derecha de cada histograma hasta obtener una sumatoria que resulte en 0,05 o más, esto hizo que para cada probabilidad condicional en cada año se obtuviese un $\alpha$ distinto donde en todos los casos estos deberían ser el menor número posible que fuese mayor a 0,05. Los resultados obtenidos se pueden ver en las figuras 3) y 4).

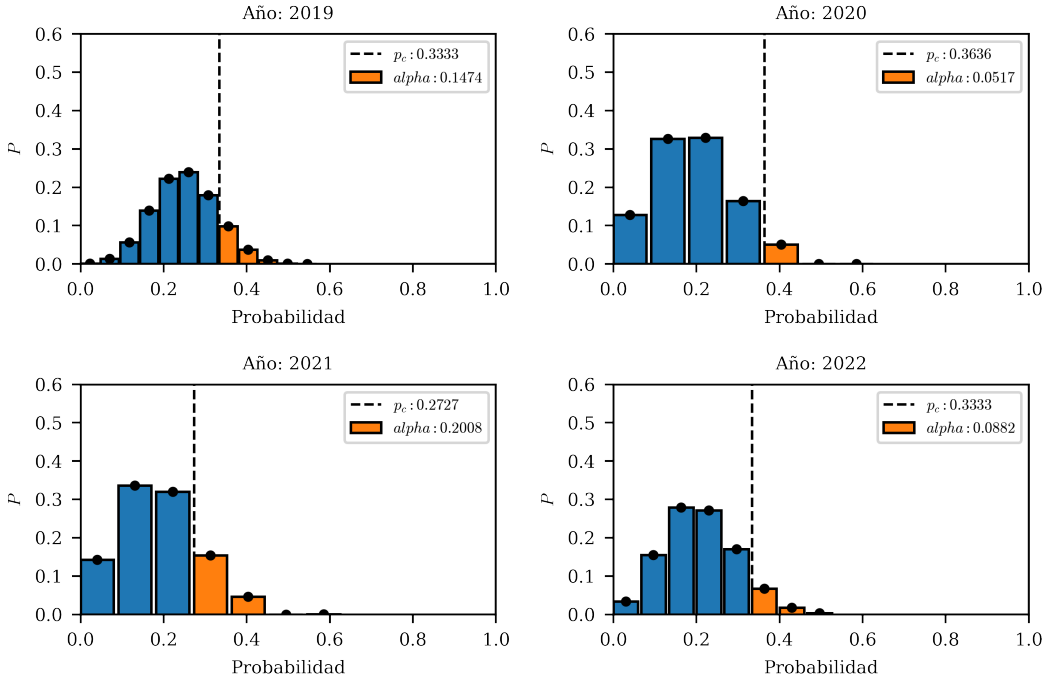



Figura 3: Distribución de la probabilidad condicional $P(M|m)$ mostrando en cada año la probabilidad crítica correspondiente con su $\alpha$.

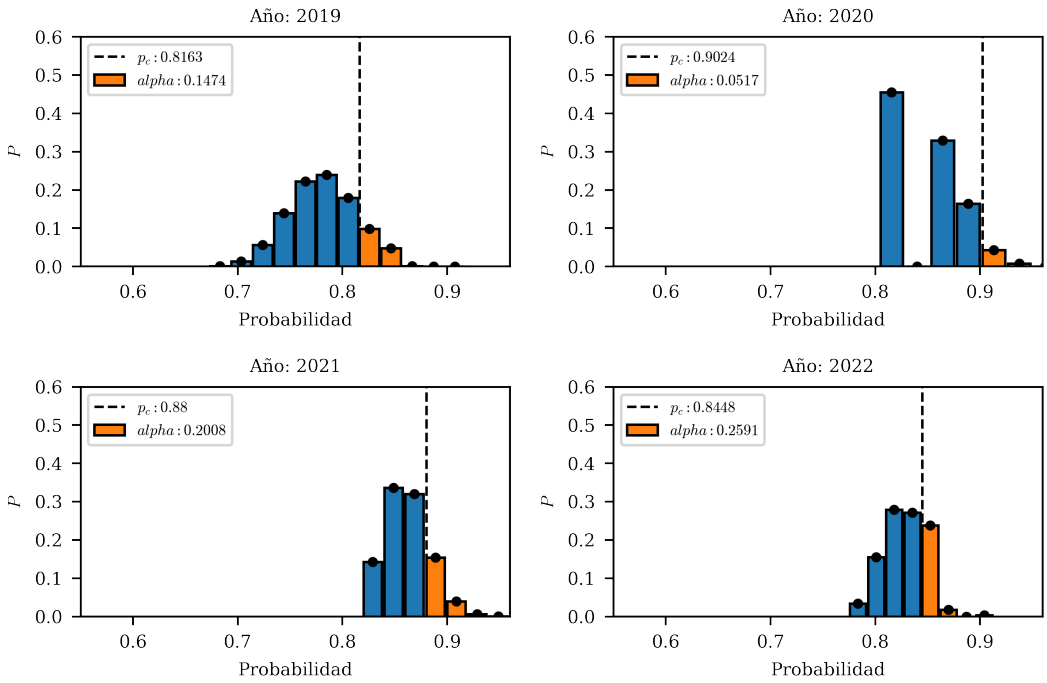



Figura 4: Distribución de la probabilidad condicional $P(H|h)$ mostrando en cada año la probabilidad crítica correspondiente con su $\alpha$.

Se puede observar que los $\alpha$'s de las probabilidades condicionales coinciden año a año, excepto en el año 2022. Esta coincidencia en el área que contemplan las zonas de rechazo muestra que la probabilidad de obtener probabilidades condicionales $P(M|m)$ y $P(H|h)$ más extremas son iguales, por lo menos en los años mencionados.

## 3. Aplicando el test

En las siguientes figuras se graficó el resultado de las distribuciones de $P(M|m)$ y $P(H|h)$ con su $P_C$ y con el valor de probabilidad condicional reportado en la Tabla 1. Se tiene entonces que si el valor de $P_C$ es mayor que la probabilidad condicional reportada no se rechaza la $H_0$, mientras que si es al revés se rechaza.

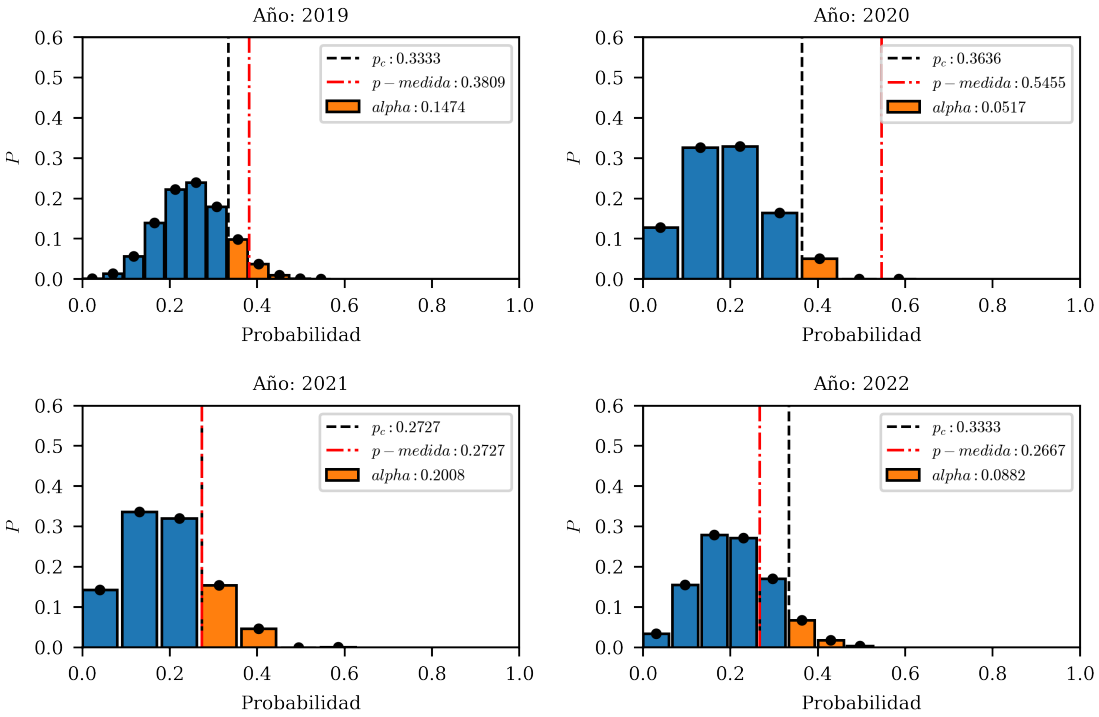



Figura 5: Distribución de la probabilidad condicional $P(M|m)$ mostrando en cada año la probabilidad crítica correspondiente con su $\alpha$ y con la probabilidad $(p-medida)$ reportada en la Tabla 1.

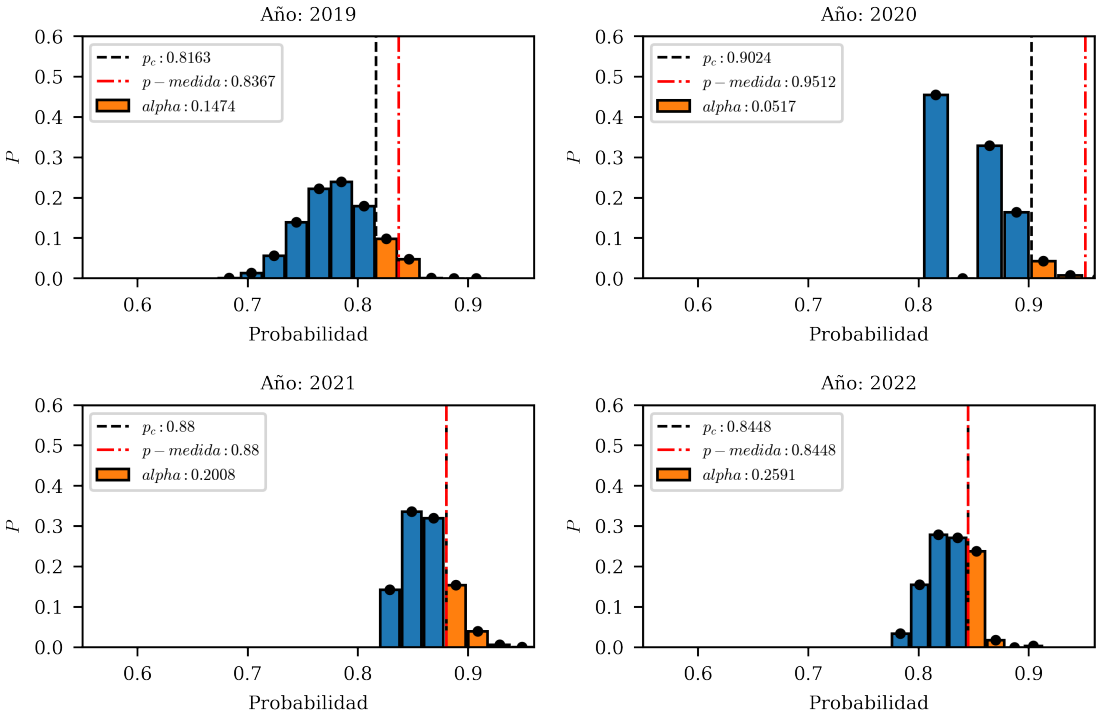



Figura 6: Distribución de la probabilidad condicional $P(H|h)$ mostrando en cada año la probabilidad crítica correspondiente con su $\alpha$ y con la probabilidad $(p-medida)$ reportada en la Tabla 1.

Análogo a lo que pasó en el resultado anterior, se observa una cierta coincidencia de los resultados anualmente. Donde en los años 2019, 2020, y 2021 el número de bines que separa la $P_C$ de la $p-medida$ es igual. Entonces, siguiendo la lógica explicada anteriormente, en los años 2019 y

2020 se rechaza la $H_0$ ya que la probabilidad condicional reportada se encuentra en la región de rechazo, mientras que en los años 2021 y 2022 no se rechaza la $H_0$. Con los valores de probabilidades condicionales reportadas luego se procedió a calcular en $p-value$ en cada caso, sumando los bines a derecha del valor reportado. Los resultados se observan en el Cuadro 1y se analizan en el siguiente punto.

## 4. Potenciando el test

Observando el Cuadro 1 se puede ver que los $p-values$ coinciden año a año. Esto es coherente ya que las probabilidades condicionales medidas $P(M|m)$ y $P(H|h)$ corresponden a una misma situación. Es decir, debido a que se tienen las mismas condiciones de contorno, si se conoce una probabilidad condicional la otra queda fijada para que cumpla con esas condiciones. Eso hace entonces que tenga sentido que el $p-value$, osea la probabilidad de obtener esa probabilidad condicional u otra más extrema, sea la misma para $P(M|m)$ y para $P(H|h)$.

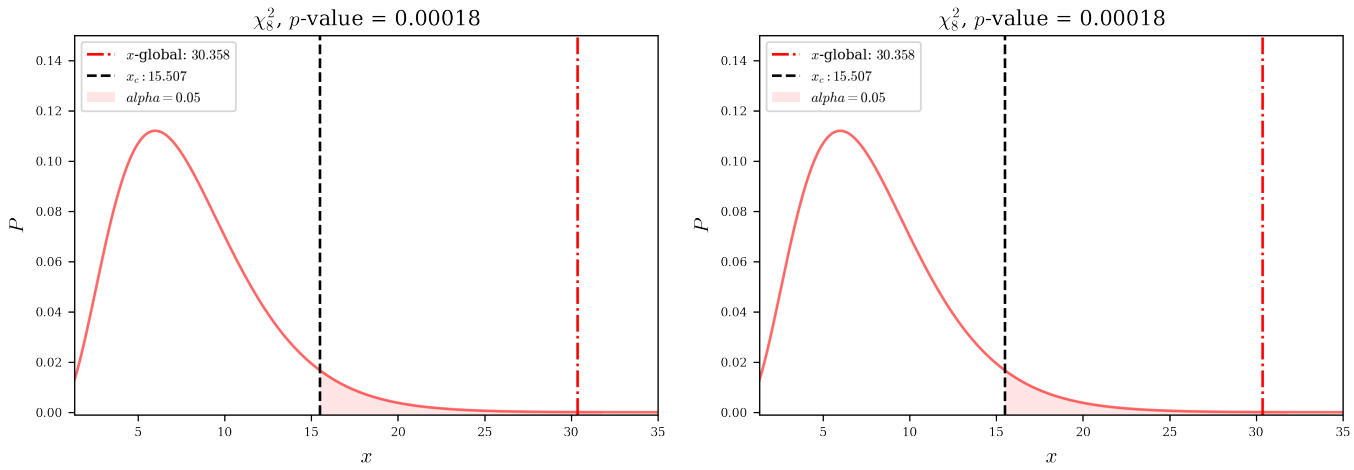
Es relevante aclarar que para el caso particular del 2020 el $p-value$ se estimó considerando el cociente $\frac{1}{N}$ debido a que a derecha de esa probabilidad condicional reportada la sumatoria daba cero, por lo que una forma de acotar el $p-value$ era considerando el número de veces que se simuló el problema descrito.

Habiendo obtenido entonces estos p-valores, se pide en este punto combinar todos mediante el estadístico $\chi_{2n} = -2ln(\prod_i^n p_i)$, donde en este caso $n = 8$ debido a que se tienen 4 resultados en cada probabilidad condicional.

| | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|
| $P(H\|h)$ | 0,0492 | 0,0001 | 0,2008 | 0,2591 |
| $P(M\|m)$ | 0,0492 | 0,0001 | 0,2008 | 0,2591 |

Cuadro 1: Tabla con el resultado de los p-valores correspondientes a las probabilidades condicionales reportadas.

Conociendo la distribución de este estadístico ($\chi_{2n}$) y obteniendo el p-valor conjunto se puede volver a analizar si se rechaza o no la $H_0$ con una significancia de $\alpha = 0{,}05$. Lo que se hizo entonces fue calcular el valor crítico mediante la integral a derecha de la distribución $\chi_{2n}$. Y de la misma forma si el p-valor conjunto es mayor que el valor crítico $H_0$ entonces la misma se rechaza, mientras que si es al revés no se rechaza. El resultado obtenido se puede ver en las siguientes figuras.




(a) (b)

Figura 7: Distribución de $\chi^2_8$ con la $x_c$ para el cual $\alpha = 0{,}05$ y mostrando explícitamente el valor de la distribución en el que está el resultado de combinar los $p - values$ obtenidos.

Se puede observar que debido que para ambas probabilidades condicionales los $p-values$ fueron iguales entonces el resultado de la fórmula descrita y el $p-value$ obtenido es el mismo. Y debido a que en ambos casos el $x_c$ para el que $\alpha = 0{,}05$ está a la izquierda del $x - global$ obtenido entonces la hipótesis nula $H_0$ se rechaza sobre la totalidad de la tabla, es decir, observando globalmente los resultados de la Tabla 1 se puede rechazar la hipótesis que no hay afinidad de género.

## 5. Test de Contingencia

En este punto se aplica el test de Fisher. Se procedió entonces a calcular la probabilidad de todas las posibles tablas sujetas a las condiciones de contorno impuestas en la Tabla 1. Esta probabilidad se calculó asumiendo $H_0$ como verdadera, entonces las posibles tablas cuentan con una distribución hipergeométrica. Para calcular el $p - value$ con el Test de Fisher en cada año fue necesario conocer a qué tabla correspondían los datos reportados. Conociendo ese valor bastaba con sumar la probabilidad de obtener esa tabla y tablas más 'extremas', es decir, aquellas que se acercaban a casos de alta afinidad de género, y así de esa sumatoria se obtendría el $p - value$. Usando las probabilidades condicionales reportadas se calculó para cada año cual fue la tabla obtenida y se procedió a obtener el $p - value$ correspondiente. Los resultados se observan en el Cuadro 2.

| | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|
| $p - value$ | 0,01247 | 0,00002 | 0,04700 | 0,08755 |

Cuadro 2: Tabla con el resultado de los p-valores obtenidos usando el test de Fisher.

En primer lugar se observa que los $p - values$ del Cuadro 1 y del Cuadro 2 no coinciden. Sin embargo, si se conserva el orden creciente de los $p - values$ siendo el menor de todos el del año 2020, el siguiente el del 2019, el siguiente el del 2021, y finalmente el mayor el del 2022. Por otro lado, en este caso $H_0$ se rechaza en el 2019, 2020 y 2021, ya que para estos valores $p - value < \alpha = 0{,}05$, aunque como se puede ver, el $p - value$ del 2021 es el que se encuentra más

cercano al $\alpha$ establecido difiriendo en menos del 6 % del valor obtenido, lo que muestra una cierta concordancia con los resultados de rechazo de $H_0$ vistos con el test anterior, donde en el 2021 no se rechaza la $H_0$. A su vez, es relevante mencionar que en este Test, a diferencia del anterior, el $\alpha$ es fijo en todos los casos, mientras que, como se ve en las Figuras 5 y 6, los $\alpha$'s considerados podían ser mayores a 0,05 debido la postura conservadora adoptada. Viendo explícitamente el caso del 2021 se tenía $\alpha = 0{,}2008$ por lo que habría sido rechazada $H_0$ en este caso si se hubiese podido considerar un $\alpha = 0{,}05$ .Finalmente podemos hacer uso una vez más del estadístico definido en el punto 4 y estudiar el resultado global obtenido.

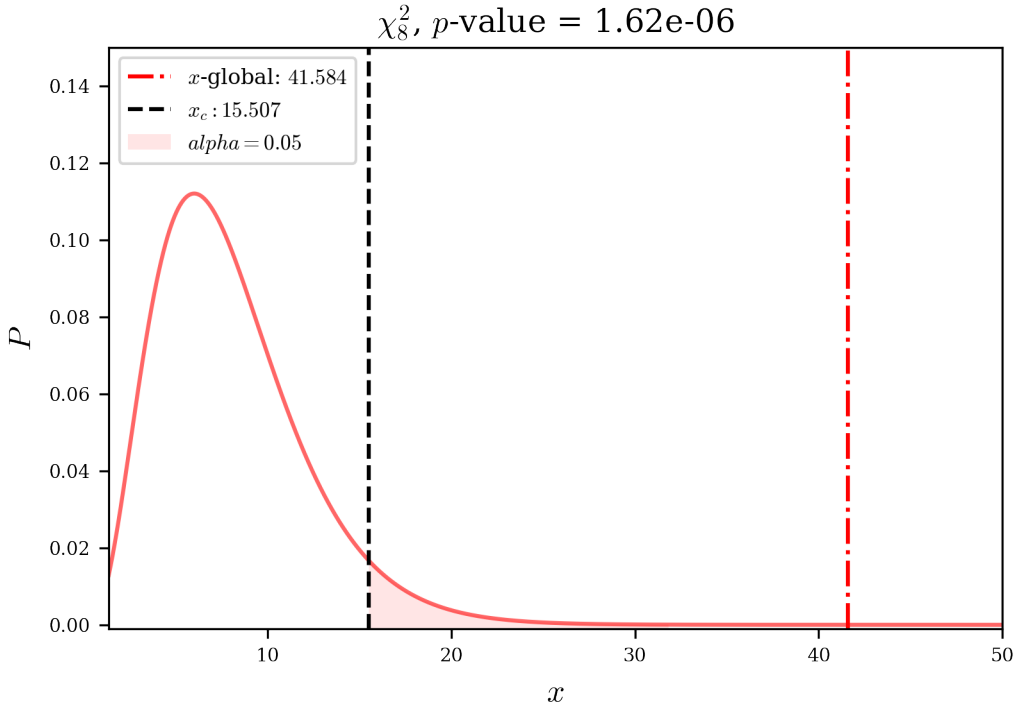



Figura 8: Distribución de $\chi^2_8$ con la $x_c$ para el cual $\alpha = 0{,}05$ y mostrando explícitamente el valor de la distribución en el que está el resultado de combinar los $p - values$ obtenidos.

Se puede observar que en este caso el $p - value$ resultante de la combinación de $p - values$ es menor por dos ordenes de magnitud al visto anteriormente en la figura 7. De todas formas, en ambos casos, globalmente se rechaza $H_0$ debido a que el $x - global$ está en la zona de rechazo, definida con una significancia de 0,05.

## 6. Conclusión

Se observó entonces que los $p - values$ eran independientes de la probabilidad condicional considerada, sino que quedaban definidos anualmente. Por otra parte, calcularon distintos $p-values$ para el primer y segundo test, conservando anualmente su orden creciente. Se observó que el caso que mostraba resultados más coincidentes con $H_0$ fue el del 2022, mientras que donde se observó mayor afinidad de género fue en el 2020. Por otro lado, en el primer test $H_0$ se rechazó en el 2020 y 2019, mientras que en el segundo test se rechazó $H_0$ en el 2021, 2020 y 2019. Sin embargo, el $p-value$ reportado para el 2021 contaba con una diferencia del 6 % con la significancia establecida de 0,05. Finalmente, en ambos tests estudiando los datos obtenidos globalmente se rechazó $H_0$.