¿Existe diferencia entre el consumo de cerveza entre las personas que practican deporte y las que no?

Autor:

AdriCri22

Índice

Resumen

Objetivos: Ver si la gente que practica alguna actividad física consume la misma cantidad de cerveza que la que no practica deporte.

Recogida y tratamiento de los microdatos: Hemos obtenido los metadatos del INE, después hemos tratado los datos para obtener las variables que nos interesan con R.

Comparación de poblaciones: Nuestros datos son independientes, no siguen una distribución normal pero como el número de observaciones es mayor a 100 aplicamos una distribución normal. Calculamos el estadístico y vemos que podemos rechazar nuestra hipótesis nula, concluyendo que la gente que practica deporte no bebe la misma cantidad de cerveza.

Previsión: Al analizar nuestros datos en los gráficos obtenidos en R, observamos que no siguen las premisas de independencia, normalidad y homocedasticidad, pero sí la de linealidad, por ello al calcular cualquier previsión nuestros resultados no serán muy fiables.

Objetivos

El consumo de alcohol ha estado presente en la sociedad desde hace más de 10 millones de años, generando un gran conjunto de mitos y creencias como, por ejemplo, su uso medicinal o estimulante.

Generalmente, al pensar en alcohol, lo primero que se nos pasa por la cabeza es la cerveza. Inicialmente desarrollada en pueblos egipcios y sumerios, hoy día es consumida en gran parte del mundo en alguna de sus muchas variedades, que van desde los diferentes tipos de cereales usados hasta los diferentes tipos de aditivos aromáticos.

Por ello, se considera que la cerveza es la bebida alcohólica más consumida del mundo. No obstante, hay casos en los que esta está relacionada con malos hábitos, excesos y adicciones. Consecuentemente, es importante consumirla con moderación y seguir una vida sana que, a grandes rasgos, se basa en una dieta equilibrada y en el deporte.

Nos hemos preguntado cuál es la relación entre el consumo de cerveza entre los diferentes grupos de población deportista y no deportistas, clasificándolos según diversos filtros, como pueden ser la edad, el sexo, el peso y la altura, las horas de ejercicio diarias, etc. Por otro lado, hemos querido hacer una predicción sobre cuál es el impacto que tiene la gente que hace deporte en el consumo de cerveza y además hemos comparado el consumo de cerveza es el mismo entre mujeres y hombres, fumadores y no fumadores, andaluces y catalanes y hemos analizado por encima los gráficos del consumo de cerveza en relación a la edad, el peso y la altura.

Recogida de datos

El proceso de recogida de datos está dividido en dos partes:

Obtención de datos

Los datos han sido obtenidos de la página del INE (Instituto Nacional de Estadística) https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176783&menu=resultados&secc=1254736195650&idp=1254735573175#!tabs-1254736195295, donde a partir de los microdatos que nos ofrecen en el apartado de salud, podemos ver que la cantidad es de 23089 personas entrevistadas y 456 variables.

Nosotros hemos seleccionado las variables que nos interesan que son:

Variables	Descripción
CCAA	La comunidad autónoma en la que reside
SEXOa	Su género
EDADa	Su longevidad
S109	Su altura
S110	Su peso
T114_1 y T114_2	Tiempo dedicado a actividades físicas intensas (horas y minutos)
T116_1 y T116_2	Tiempo dedicado a actividades físicas moderadas (horas y minutos)
V122	Tipo de tabaco que fuma (Para poder ver si fuman o no)
Desde W128Cer_1 hasta W128Cer_7	La cantidad de cerveza que consume cada día de la semana

Procesamiento de los datos

Una vez obtenidos los datos vamos a procesarlos exportando el archivo xlsx a R para poder tratarlos. Después dividiremos las variables en numéricas y categóricas y posteriormente nos quedaremos con una muestra lo más real posible, por lo que tratamos las variables reduciendo el número de la muestra a aquellas personas que hayan contestado a todas las preguntas de variables numéricas, ya que si no estas contarían como si hubieran contestado 0, dañando la muestra.

Según la información que nos dan (variables desde W128Cer_1 hasta W128Cer_7), el consumo de cerveza corresponde a un botellín de 333 ml por lo que para obtener el alcohol en litros multiplicamos la cantidad dada por 0,333 y sumamos los 7 días de la semana de consumo de esta bebida alcohólica. Para calcular el deporte que hace sumamos la cantidad de deporte intenso y moderado, las horas y minutos, convirtiendo el tiempo de deporte en una sola variable.

Dado que la muestra restante es relativamente grande, el programa elegirá aleatoriamente un subconjunto de 5 mil personas para que todas las entradas tengan la misma probabilidad de aparecer en la muestra, estos datos son una muestra de la población a la que queremos analizar.

Variables recogidas

Las “variables categóricas” son, de cada persona, su género, en qué comunidad autónoma reside y el tipo de tabaco que consume.

Las “variables numéricas” son, de cada persona, su edad, su peso, su estatura, el tiempo que dedica a actividades físicas intensas y moderadas, y el consumo de cerveza por día de la semana.

La edad, el peso y la estatura son variables discretas. La edad puede valer un número entero igual o superior a 15, siendo 15 años la edad mínima para poder participar en la encuesta realizada. El peso y la estatura son valores enteros positivos.

El tiempo que dedica a actividades físicas intensas y moderadas es el número de horas y minutos que dedica esa persona cada día en la realización de actividades físicas, distinguiendo entre las de mayor (las variables T114_1 y T114_2) o menor (las variables T116_1 y T116_2) intensidad.

El consumo de cerveza por dia de la semana es un número entero positivo que indica las unidades de cerveza que ha consumido esa persona, normalmente en forma de latas o botellines. No se pone en conjunto de toda la semana para saber si dependiendo del día el consumo varía, por si es fin de semana o día laboral.

Análisis estadístico

Comparación de poblaciones

En este estudio vamos a comprar la cantidad de cerveza que ingieren las personas que realizan deporte y la cantidad de los que no realizan ninguna actividad deportiva.

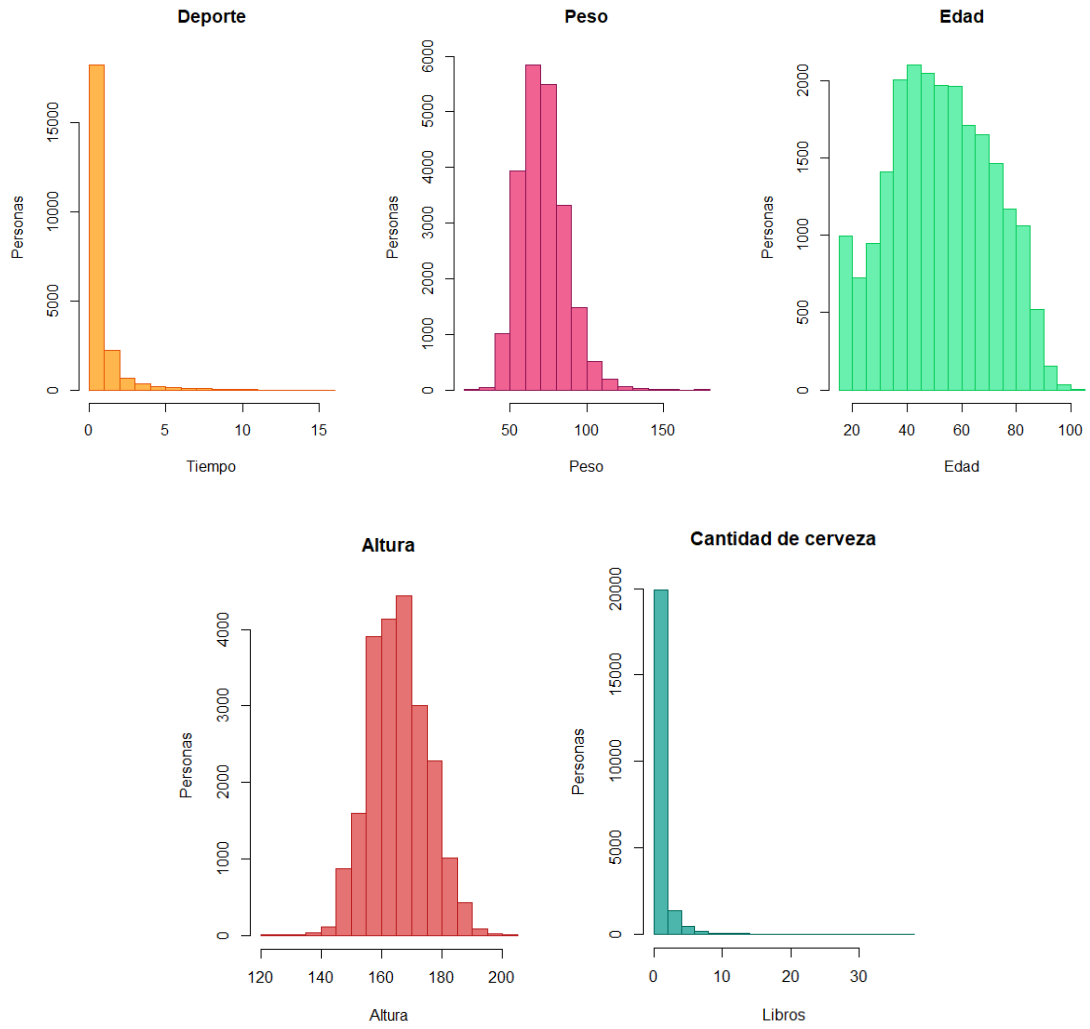
Como en este caso vamos a comprar 2 medias, nos preguntamos si realizar alguna actividad conlleva a un consumo menor o mayor de alcohol, diríamos que el consumo de alcohol depende del deporte que hagamos.

Nuestra hipótesis nula sería suponer que las dos medias (cerveza si haces deporte vs cerveza si no lo haces) valdrían lo mismo, y como hipótesis alternativa, diríamos que las medias son distintas.

$$H_0\colon\ \mu_D = \mu_{ND} \ \text{ vs } \ H_1\colon\ \mu_D \neq \mu_{ND}$$

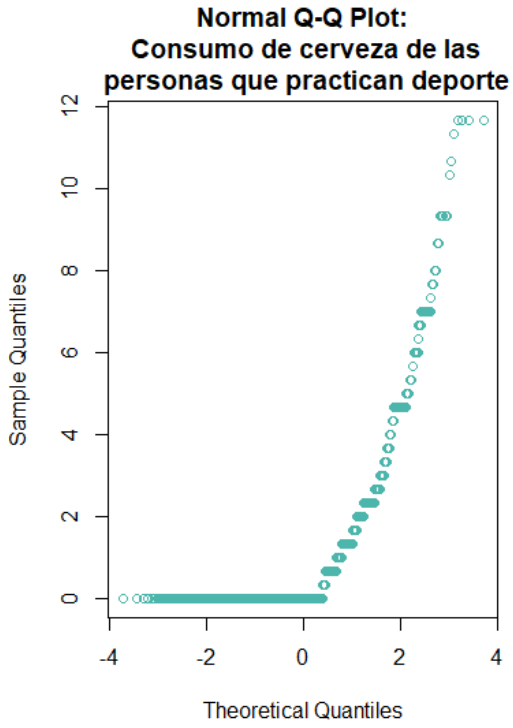
Para saber qué distribución del estadístico vamos a estudiar las premisas: en nuestro caso las variables son independientes, ya que por ejemplo una persona

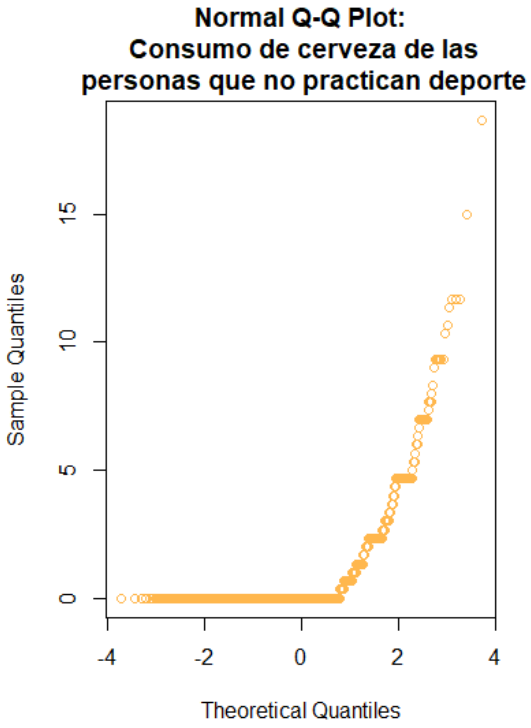
deportista no puede pertenecer al conjunto no deportista, es decir, cada caso se mide de manera independiente. Vamos a ver si nuestras variables siguen una distribución normal, para ello observamos los histogramas de las variables numéricas:

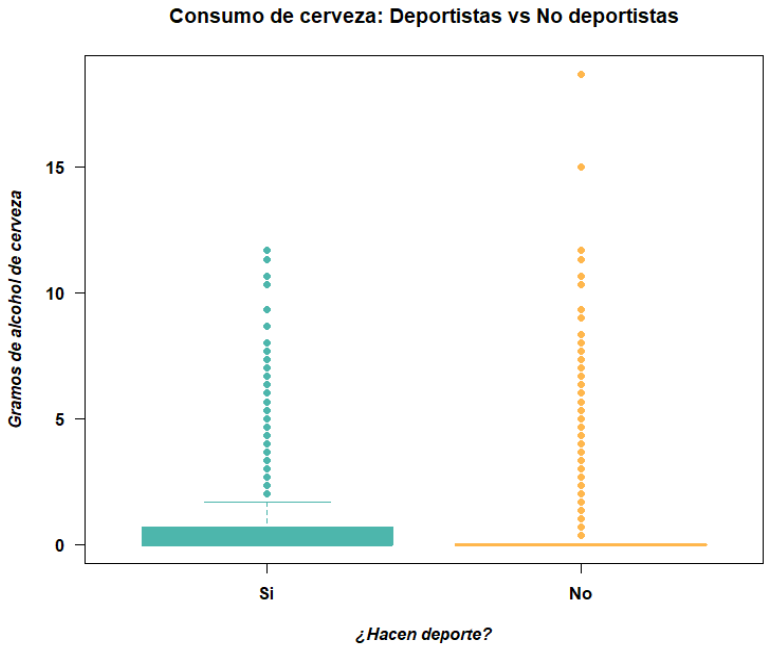


Podemos deducir que el tiempo que las personas dedican a hacer deporte y el consumo de cerveza no siguen una distribución normal, en cambio la altura, el peso y la edad de los entrevistados sí que sigue una distribución normal.

Y finalmente, podemos observar a continuación como las muestras escogidas para nuestra prueba tampoco siguen una distribución normal, esto lo podemos ver con la función qqnorm y boxplot en R:







A partir de las premisas, concluimos que el estadístico que debemos usar es el de la distribución normal, ya que el número de observaciones es mayor a 100. A continuación procedemos a calcular el estadístico, para ello explicaremos cómo hemos obtenido los parámetros.

Nota: X_D, indica que un parámetro referente a la muestra que hace deporte, X_{ND} hace referencia a la muestra que no hace deporte.

Hay que recalcar que hemos reducido el tamaño de las muestras a 5000 cada una, por lo que:

$$n_D = n_{ND} = 5000$$

La media la obtenemos mediante la función en R mean()

$\underline{Y_D} = 0.6360966$ $\underline{Y_{ND}} = 0.444222$

La desviación tipo la obtenemos mediante la función en R var()

$S_D = 1.658714$ $S_{ND} = 1.609415$

Calculamos el valor del estadístico:

$$Z = \frac{\left(\underline{Y_D} - \underline{Y_{ND}}\right)}{\sqrt{S_D/n_D + S_{ND}/n_{ND}}} = \frac{(0.6360966 - 0.444222)}{\sqrt{1.658714/5000 + 1.609415/5000}} = 7.505038$$

Calculamos el punto crítico: $\alpha = 0.05 \rightarrow z_{1-\alpha/2} = 1.959964 < 7.505038 = Z$

Conclusión:

Dado que el punto crítico es menor al estadístico podemos rechazar la hipótesis nula, no tenemos evidencias para asegurar que la media de cerveza consumida semanalmente sea igual para la gente que practica deporte y los que no practican.

Además hemos contrastado nuestras conclusiones con la función t.test de R, la cual nos confirma ya lo dicho anteriormente.

Inferencia con intervalo de confianza:

$$\sigma = s = \sqrt{\frac{(n_D - 1)\cdot S_D^2 + (n_{ND} - 1)\cdot S_{ND}^2}{(n_D - 1) + (n_{ND} - 1)}} =$$

$$= \sqrt{\frac{(5000 - 1)\cdot 1.658714^2 + (5000 - 1)\cdot 1.609415^2}{(5000 - 1) + (5000 - 1)}} = 1.63425$$

$$IC(\mu_D - \mu_{ND}, 95\%) = \left(\underline{Y_D} - \underline{Y_{ND}}\right) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n_D} + \frac{s^2}{n_{ND}}} =$$

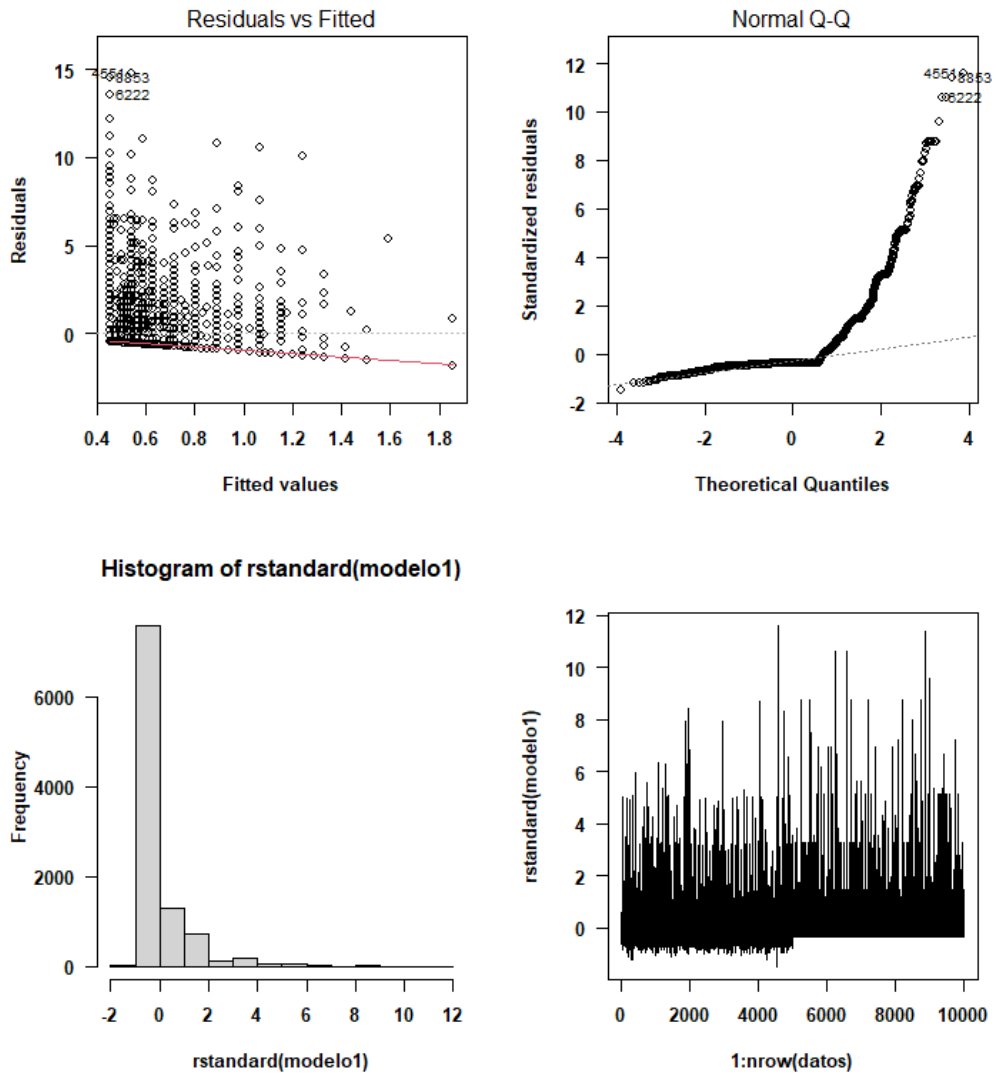
$$= (0.6360966 - 0.444222) \pm 1.959964 \cdot \sqrt{\frac{1.63425^2}{5000} + \frac{1.63425^2}{5000}}$$

$$= (0.1278132, \quad 0.2559360)$$

Podemos asegurar que con un 95% de confianza los no deportistas consumen una media de 0.1278132 y 0.2559360 litros de cerveza menos.

Previsión

Para poder hacer una previsión de cuántos litros de cerveza consumiría una persona que hace x minutos de deporte procedemos a obtener los gráficos para validar nuestro modelo lineal:



A la hora de validar el modelo lineal podemos ver que de todas las premisas sólo se cumple 1 la de linealidad, lo podemos comprobar en el gráfico de arriba a la izquierda, en este mismo gráfico podemos observar que hay heteroscedasticidad, en el gráfico de arriba a la derecha y abajo a la izquierda podemos observar que nuestros datos no siguen una distribución normal y finalmente en el gráfico de abajo a la derecha vemos que hay dependencia, porque nuestros residuos muestran cierto patrón, ya que vemos que las bajadas y subidas acaban aproximadamente en el mismo eje horizontal excepto por unos casos en los que podemos ver que unos datos se disparan siguiendo una tendencia creciente. Por lo que al hacer nuestra previsión nuestra conclusión no será muy fiable.

Si intentamos hacer una transformación logarítmica nos salen valores infinitos en nuestros datos, por lo que no podemos aplicar dicha transformación y por lo tanto, no podemos solucionar el no cumplimiento de las premisas.

Vamos a calcular los diferentes parámetros para posteriormente hacer los cálculos necesarios:

La variable condición X corresponde a el tiempo de deporte en horas.

La variable respuesta Y corresponde a la cantidad de cerveza.

$$n = n_D + n_{ND} = 5000 + 5000 = 10000$$

$\underline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = 0.5401593$ (resultado obtenido mediante la función mean() en R)

$\underline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = 0.9678683$ (resultado obtenido mediante la función mean() en R)

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}}{n-1}} = \sqrt{\frac{19347.14 - \frac{5401.593^2}{1000}}{10000-1}} = 1.281837$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}} = \sqrt{\frac{32296.81 - \frac{9678.683^2}{1000}}{10000-1}} = 1.514312$$

$$S_{XY} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{n-1} = \frac{7237.866 - \frac{5401.593 \cdot 9678.683}{10000}}{10000-1} = 0.2010036$$

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{0.2010036}{1.514312 \cdot 1.281837} = 0.1035513$$

$$S^2 = \frac{(n-1) \cdot (S_Y^2 - b_1 \cdot S_{XY})}{n-2} = \frac{(1000-1) \cdot (1.281837^2 - 0.08765427 \cdot 0.2010036)}{1000-2} = 1.62565$$

Calculamos los estimadores:

$$b_1 = \frac{S_{XY}}{S_X^2} = \frac{0.2010036}{1.514312^2} = 0.08765427$$

$$b_0 = \underline{Y} - b_1 \cdot \underline{X} = 0.5401593 - 0.08765427 \cdot 0.9678683 = 0.4553215$$

Calculamos las varianzas:

$$S_{b_0} = S^2 \left(\frac{1}{n} + \frac{\underline{X}}{(n-1) \cdot S_X^2} \right) = 1.62565 \cdot \left(\frac{1}{10000} + \frac{0.9678683}{(10000-1) \cdot 1.514312} \right) = 0.01513211$$

$$S_{b_1} = \frac{S^2}{(n-1) \cdot S_X^2} = \frac{1.62565}{(10000-1) \cdot 1.514312} = 0.008420149$$

Vamos a predecir cuántas cervezas tomaría una persona que hace 7 horas de deporte a la semana:

Estimación puntual:

$\hat{y}_h = b_0 + b_1 \cdot x_h = 0.4553215 + 0.08765427 \cdot 7 = 1.0689$ cervezas a la semana

Intervalo de confianza para el valor esperado esperado:

$t_{n-2,0.975} = 1.960201$ (calculado mediante el código: pt(0.975, 9998))

$$IC = \hat{y}_h \pm t_{n-2,0.975} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x_h - \underline{X})^2}{\sum (X_i - \underline{X})^2}} =$$

$$= 1.068901 \pm 1.960201 \cdot 1.27501 \cdot \sqrt{\frac{1}{10000} + \frac{(7 - 0.9678683)^2}{-5.089262 \cdot 10^{-13}}} =$$

$$= (0.9662509, \quad 1.1715519)$$

Ahora queremos contrastar si podemos aceptar que el tiempo de deporte se incrementa en 1 h por cada cerveza de más que la persona bebe.

Hipótesis:

$$H_0: \beta_1 = 1 \quad H_1: \beta_1 \neq 1$$

Calculamos el estadístico:

$$t_{b_1} = \frac{(b_1 - \beta_1)}{S_{b_1}} = \frac{(0.08765427 - 1)}{0.008420149} = -108.3527$$

Calculamos el p-valor:

$t_{n-2,0.975} = 1.960201$ (Calculado anteriormente)

Dado que $|t_{b_1}| > t_{n-2,0.975}$, no es verosímil que el coeficiente de la pendiente sea 1, como ya hemos dicho anteriormente, nuestras predicciones no serán correctas dado que los datos obtenidos no cumplen ninguna de las premisas del modelo lineal.

summary() del modelo señalando los parámetros usados:

```
Call:
lm(formula = datos$ConsumoCerveza ~ datos$TiempoDeporte, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8578 -0.5430 -0.4553 -0.1223 14.7750

Coefficients:
                    Estimate Std.      Error       t value           Pr(>|t|)
(Intercept)          b0 = 0.45532   Sb0 = 0.01513   tb0 = 30.09   p − valorb0 = <2e-16
***
datos$TiempoDeporte  b1 = 0.08765   Sb1 = 0.00842   tb1 = 10.41   p − valorb1 = <2e-16
***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: S =1.275 on grados de libertad = 9998 degrees of freedom
Multiple R-squared:  R2 = rXY2 = 0.01072,   Adjusted R-squared:  0.01062
F-statistic: 108.4 on 1 and 9998 DF,  p-value: < 2.2e-16
```

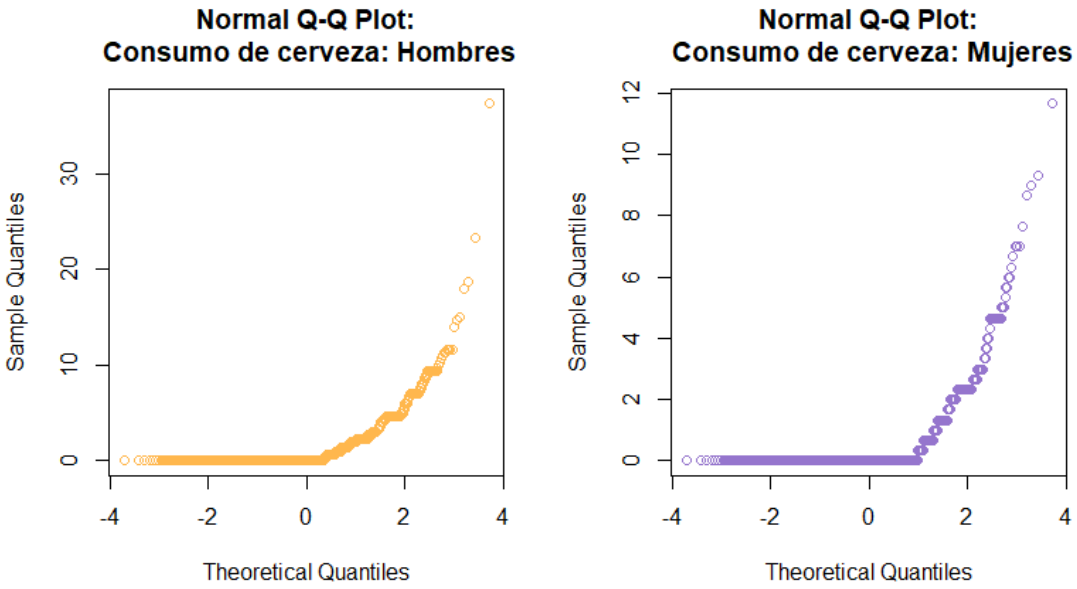
Más comparaciones

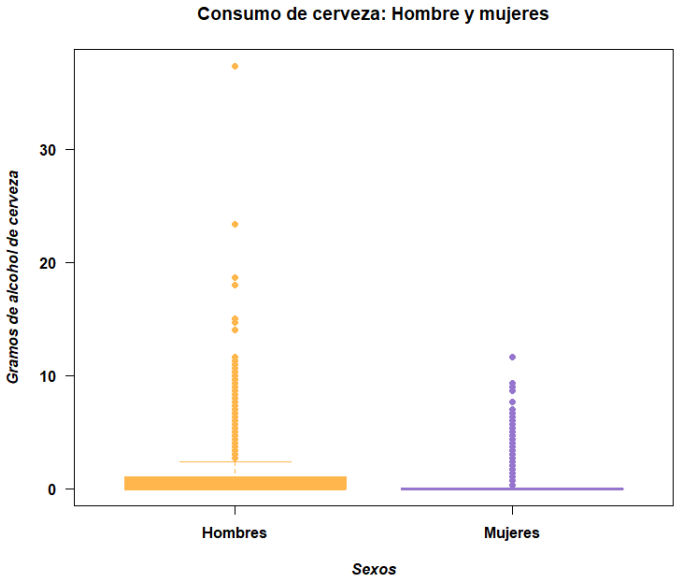
Como estudios secundarios vamos a comparar lo que siempre hemos escuchado, si los hombres consumen más cerveza que las mujeres.

Nuestra hipótesis nula sería suponer que las dos medias son iguales, y como hipótesis alternativa, diremos que los hombres consumen más cerveza que las mujeres.

$$H_0\colon\ \mu_M = \mu_H \ \text{ vs } \ H_1\colon\ \mu_M < \mu_H$$

Aquí vemos los gráficos q-q plot y boxplot:





Al estudiar las premisas, concluimos que el estadístico que debemos usar es el de la distribución normal, ya que el número de observaciones es mayor a 100.

A continuación procedemos a calcular el estadístico, para ello explicaremos cómo hemos obtenido los parámetros.

Nota: X_H, indica que un parámetro referente a la muestra con género igual a hombre, X_M hace referencia a la muestra con género igual a mujer.

Hay que recalcar que hemos reducido el tamaño de las muestras a 5000 cada una, por lo que:

$$n_H = n_M = 5000$$

La media la obtenemos mediante la función en R mean()

$$\underline{Y}_M = 0.238428 \qquad \underline{Y}_H = 0.8513478$$

La desviación tipo la obtenemos mediante la función en R var()

$$S_M = 0.5566126 \qquad S_H = 3.109018$$

Calculamos el valor del estadístico:

$$Z = \frac{\left(\underline{Y}_M - \underline{Y}_H\right)}{\sqrt{S_M/n_M + S_H/n_H}} = \frac{(0.238428 - 0.8513478)}{\sqrt{0.5566126/5000 + 3.109018/5000}} = -22.63676$$

Calculamos el p-valor (unilateral):

$$P(z > -22.63676) \approx 0 < 0.05$$

Conclusión

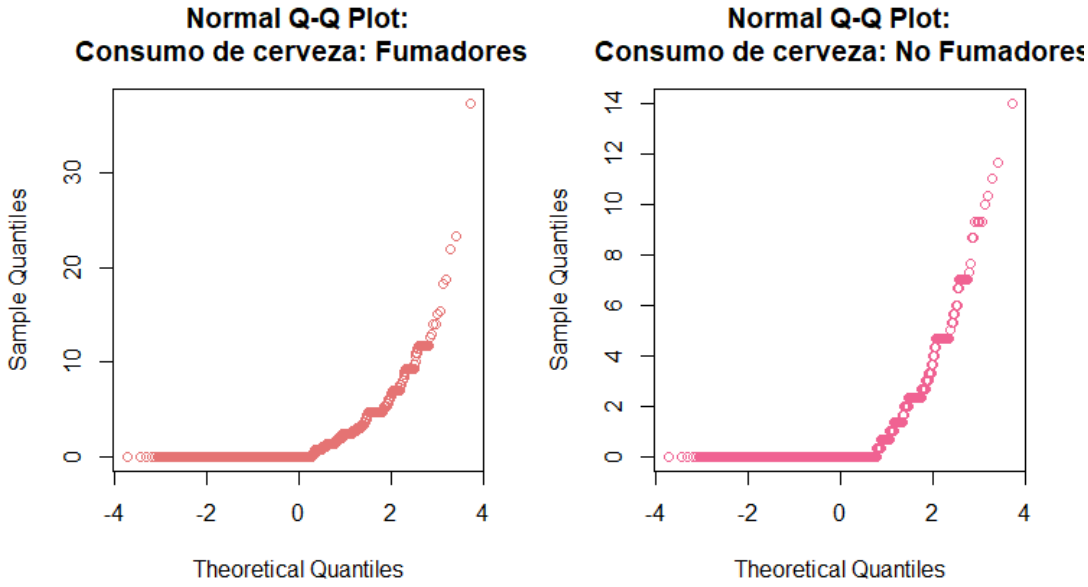
Dado que el p-valor es menor que alfa (0.05) podemos rechazar la hipótesis nula, tenemos evidencias de que los hombres consumen más cerveza que las mujeres.

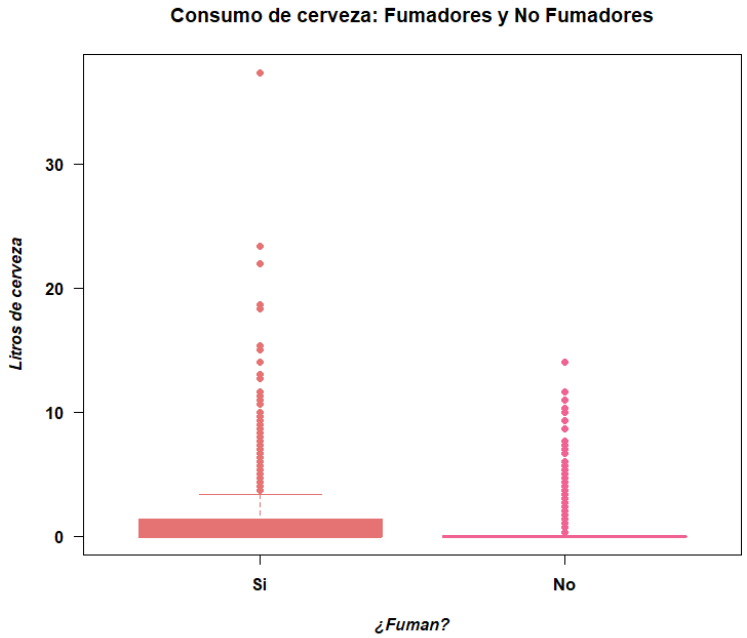
Otra prueba que hemos querido realizar es si las personas que fuman consumen más cantidad de cerveza que las que no fuman.

Nuestra hipótesis nula sería suponer que las dos medias son iguales, y como hipótesis alternativa, diremos que los fumadores beben más que los no fumadores.

$$H_0\colon\ \mu_{NF} = \mu_F \ \text{ vs } \ H_1\colon\ \mu_{NF} < \mu_F$$

Aquí vemos los gráficos q-q plot y boxplot:





Al estudiar las premisas, concluimos que el estadístico que debemos usar es el de la distribución normal, ya que el número de observaciones es mayor a 100. A continuación procedemos a calcular el estadístico, para ello explicaremos cómo hemos obtenido los parámetros.

Nota: X_F, indica que un parámetro referente a la muestra fumadora, X_{NF} hace referencia a la muestra no fumadora.

Hay que recalcar que hemos reducido el tamaño de las muestras a 5000 cada una, por lo que:

$$n_F = n_{NF} = 5000$$

La media la obtenemos mediante la función en R mean()

$$\underline{Y}_{NF} = 0.3841488 \qquad \underline{Y}_F = 0.9501822$$

La desviación tipo la obtenemos mediante la función en R var()

$$S_{NF} = 1.112839 \qquad S_F = 3.644709$$

Calculamos el valor del estadístico:

$$Z = \frac{\left(\underline{Y}_{NF} - \underline{Y}_F\right)}{\sqrt{S_{NF}/n_{NF} + S_F/n_F}} = \frac{(0.3841488 - 0.9501822)}{\sqrt{1.112839/5000 + 3.644709/5000}} = -18.34997$$

Calculamos el p-valor (unilateral):

$$P(z > -22.63676) \approx 0 < 0.05$$

Conclusión:

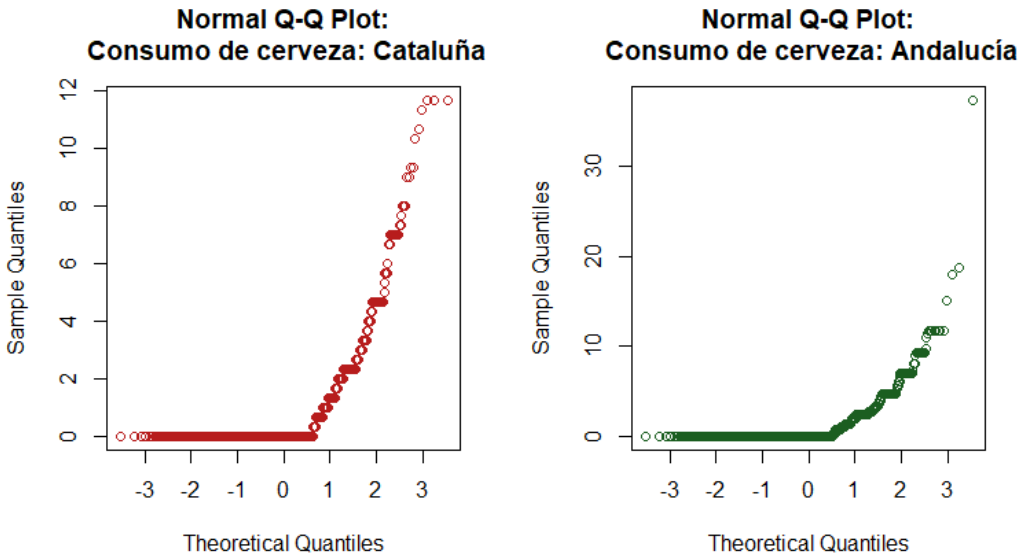
Dado que el p-valor es menor que alfa (0.05) podemos rechazar la hipótesis nula, tenemos evidencias para afirmar que las personas que fuman consumen más cerveza que las que no.

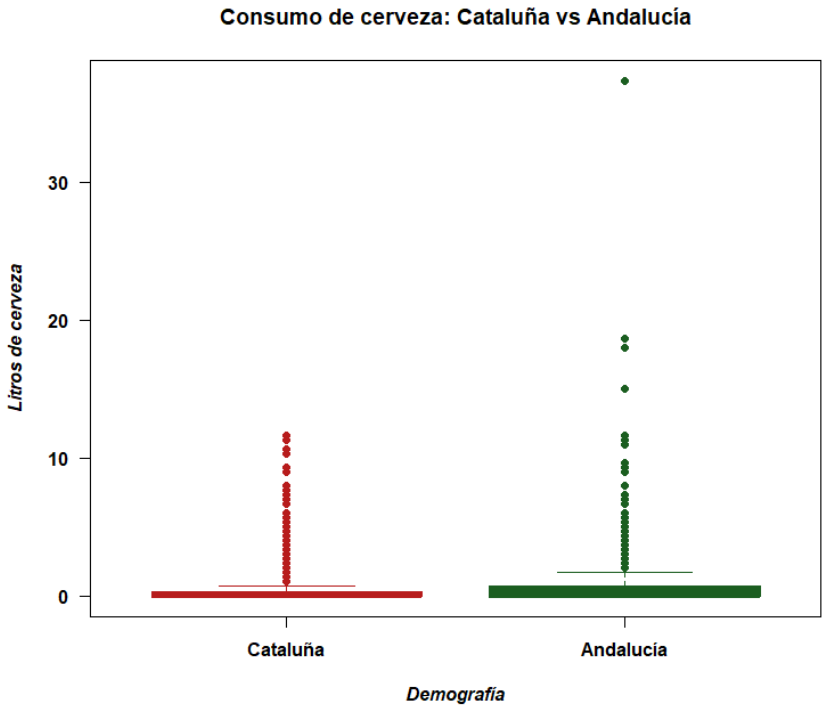
Siempre se ha dicho que al sur la población es más propensa a la fiesta y a beber, por ello hemos comparado si en Andalucía la gente bebe más cerveza que en Cataluña.

Nuestra hipótesis nula sería suponer que las dos medias son iguales, y como hipótesis alternativa, diremos que los Andaluces beben más que los Catalanes.

$$H_0\colon\ \mu_C = \mu_A \ \text{ vs } \ H_1\colon\ \mu_C < \mu_A$$

Aquí vemos los gráficos q-q plot y boxplot:





Al estudiar las premisas, concluimos que el estadístico que debemos usar es el de la distribución normal, ya que el número de observaciones es mayor a 100. A

continuación procedemos a calcular el estadístico, para ello explicaremos cómo hemos obtenido los parámetros.

Nota: X_C, indica que un parámetro referente a la muestra residente en Cataluña, X_A hace referencia a la muestra residente en Andalucía.

Esta vez hemos reducido las muestras a 2500, por lo que:

$$n_C = n_A = 2500$$

La media la obtenemos mediante la función en R mean()

$$\underline{Y_C} = 0.546786 \qquad \underline{Y_A} = 0.8095896$$

La desviación tipo la obtenemos mediante la función en R var()

$$S_C = 1.715145 \qquad S_A = 3.643547$$

Calculamos el valor del estadístico:

$$Z = \frac{\left(\underline{Y_C} - \underline{Y_A}\right)}{\sqrt{S_C/n_C + S_A/n_A}} = \frac{(0.546786 - 0.8095896)}{\sqrt{1.715145/2500 + 3.643547/2500}} = -5.676386$$

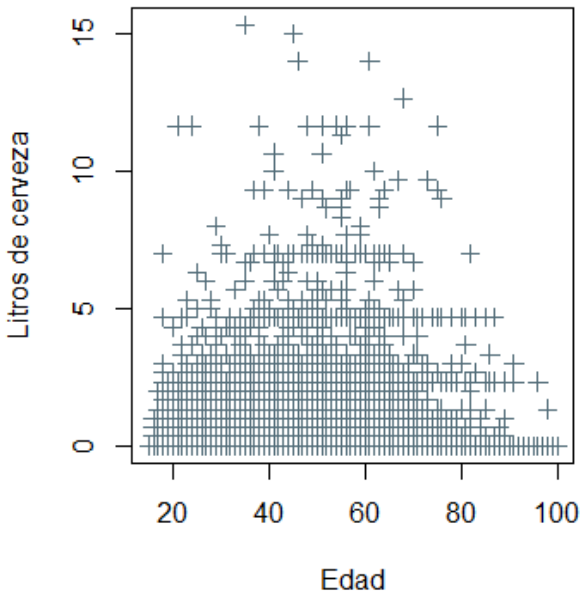
Calculamos el p-valor (unilateral):

$$P(z > -5.676386) \approx 0 < 0.05$$

Conclusión:

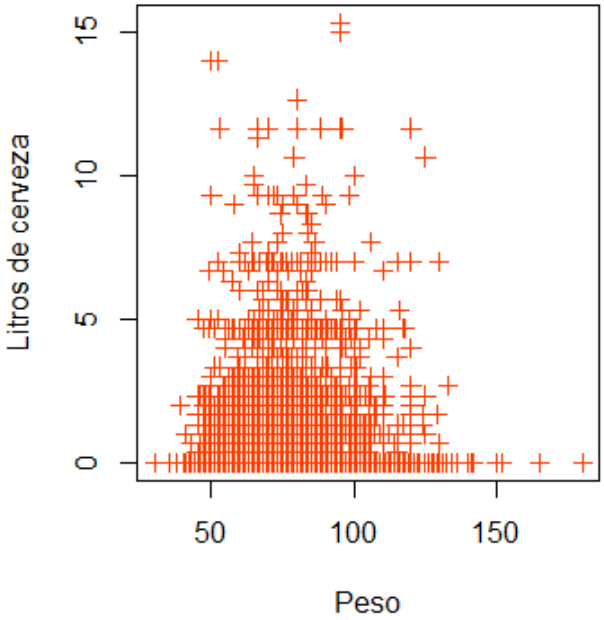
Dado que el p-valor es menor que alfa (0.05) podemos rechazar la hipótesis nula, tenemos evidencias para afirmar que los Andaluces consumen más cerveza que los Catalanes.

Al relacionar el **consumo de cerveza** con la **edad** podemos observar este gráfico:



Podemos ver que el máximo consumo de cerveza se encuentra entre los 40 y 60 años. Además entre los 15 y los 20 años, vemos un incremento del consumo de esta bebida muy abrupto, también nos podría dar a entender que esta sustancia podría empezar a consumirse incluso antes de los 15. Después de nuestro rango máximo vemos como la caída del consumo de la cerveza es más gradual.

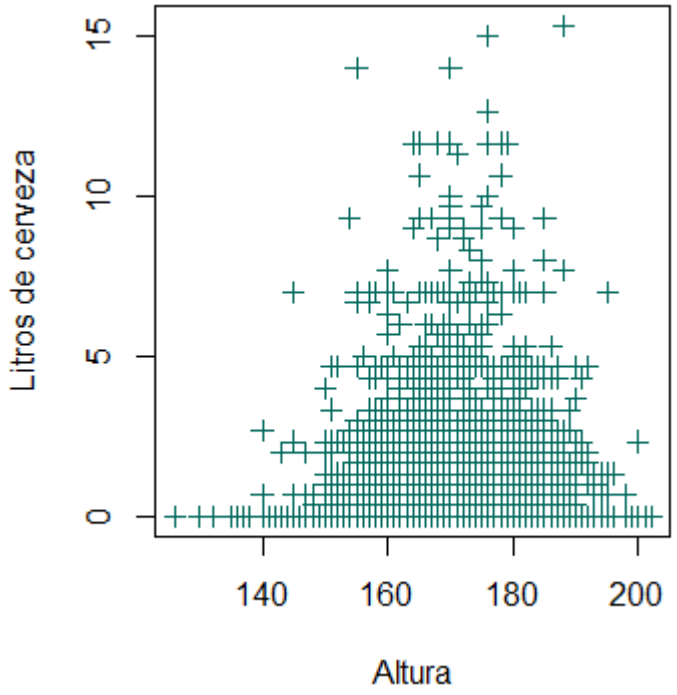
Para observar la relación entre el **peso** y el **consumo de cerveza**, tenemos:



Dónde observamos que las personas que más alcohol ingieren se encuentra entre los 50 y los 100 kilos, realmente no muestra nada relevante, ya que la

mayoría de población adulta se encuentra aproximadamente en ese rango de peso.

Por último, tenemos un gráfico que nos muestra la relación entre la altura y el consumo de cerveza:



Este gráfico tiene una curiosa forma de triángulo que nos muestra que el mayor consumo de cerveza en España se encuentra en las personas que miden 170 centímetros.