SESIÓN 2: INFERENCIA ESTADÍSTICA

1. Enumere cuales son les elementos clave de una prueba de hipótesis y su significado.

   Los elementos clave de una prueba de hipótesis son la **hipótesis nula** (H_0) que es nuestra hipótesis por defecto, la **hipótesis alternativa** (H_1) que es la hipótesis que nos interesa ya que es la variación plausible de la hipótesis nula que es la que queremos validar, el **Test statistic** que es un valor estadístico a partir de los datos muestrales que nos servirá para comparar la hipótesis nula respecto de la alternativa, la **distribución de referencia** que es la distribución del test statistic si la hipótesis es nula y por último el **p.value** que es la probabilidad de que el test statistic observado si H_0 es verdadero, si es muy alto significa que el valor statistic es normal para la distribución, por lo tanto, mi t-statistic no contradice la hipótesis nula.

2. Un elemento clave en toda prueba de hipótesis es la distribución de referencia del estadístico de la prueba. Diga de que maneras podemos conocer (o aproximar) la distribución de referencia para la prueba de hipótesis de querer ver si una media de una muestra es igual o no a un cierto valor nominal. Especifique como obtendría o cual sería esta distribución de referencia.

   Las distribuciones de referencia se construyen a partir de conjuntos de datos que, si son suficientemente amplios, permiten hacer afirmaciones sobre las probabilidades de observar determinados valores de una variable.

   Podemos conocer la distribución de referencia de dos formas:
   -**Distribución de referencia con datos historicos:**

   Primero obtendriamos la media total de los datos que queremos, seguidamente calcularíamos la distribución de referencia mediante la sieguiente formula.

   ```
   sum( value < totalMean)/n
   ```

   Donde 'n' es el número de muestras que tenemos, 'totalMean' la media total de nuestra muestra y 'value' es la medida de la muestra.

   -**Distribución de referencia sin datos historicos:**

   Primero calcularíamos la media de lamuestra y su desviación estandard, para despues poder aplicar las formulas

   ```
   t <- sqrt(n)*(totalMean - mean(x))/sd(x)
   ```

   ```
   pt(t, df)
   ```

   La función pt devuelve el valor de la función de densidad acumulada de la distribución 't' de Student dada una determinada variable aleatoria 'x' y grados de libertad 'df'. En pocas palabras, pt devuelve el área a la izquierda de un valor x dado en la distribución 't' de Student.

3. Un amigo me ofrece un piso por valor de 8MPts. Dejando de lado todos los otros factores intervinientes en la decisión de compra de un piso y solo teniendo en cuenta su precio, puedo considerar que se trata de una buena ocasión para comprar?. Que suposiciones necesito hacer para resolver el problema. (Para resolver esta pregunta utilice los datos utilizados en la clase, disponibles en el fichero: preu_3hab.r).

   Para decidir si 8MPts es una oportunidad o no, necesitamos una Distribución de Referencia de precios. Ayudándonos del fichero preu_3hab.r podemos tener bastantes datos sobre los precios de los pisos, por lo tanto, podríamos utilizar estos valores como Distribución de Referencia.

   Queda comparar todos los precios existentes con nuestro valor de 8MPts, es decir, nos interesa si es menor o igual, una vez comparados lo dividimos entre el total de muestras. Con esto tendríamos el p. value, este representa la probabilidad de que el valor 8 de la distribución de referencia de los precios históricos sea igual o menor.
   El resultado final es de un p.value **2.26%** de las veces podríamos esperar encontrar un piso con este precio o menor. Por lo tanto, podríamos decir que es una buena ocasión para comprar, ya que está por debajo del 0.05 y aceptaríamos la hipótesis alternativa.

```
p_val <- sum(x0<=8)/length(x0)
```

4. Sabemos que la media de los precios de los pisos (de 3 habitaciones) en l'Eixample es de (μ=) `16.81` y su desviación tipo es de (σ=) `5.91`. Suponiendo que la muestra obtenida en el ejercicio 1 `(7.80, 12.60, 15.96, 13.50, 8.25, 31.29, 16.46)` es aleatoria. ¿Podemos asegurar de que se trata de una muestra de pisos de l'Eixample?

   Declaramos las hipótesis nula y alternativa, $H_0 : \mu_0 \neq \mu_m$ y $H_1 : \mu_0 = \mu_m$ haciendo una comparación de las medias muestral y poblacional. En nuestro caso la media muestral es de 15.12 y la poblacional de 16.81, por lo tanto, podemos ver que las medias no son iguales, dando pie a aprobar la hipótesis nula.

   Seguidamente aplicamos una prueba de t de Student para calcular el valor probabilístico y ver el grado de confianza para esta prueba.

```
t <- sqrt(7)*(16.81 - mean(x))/sd(x)
pt(t, df=6)
```

   obteniendo una probabilidad de 0.70, al ser un valor bastante alto podemos asegurar al 70% que no se trata de una muestra de pisos del Eixample.

5. El siguiente año, los precios de una muestra aleatoria de pisos de 3hab. en l'Eixample han sido `13.57 14.80 22.36 29.29 22.70`. Puedo afirmar de que no ha habido cambio de precio entre los dos años?

```
s_pool <- sqrt(((nA-1)*var_A+(nB-1)*var_B)/(nA+nB-2))
s_pool

t <- (mitj_B - mitj_A)/(s_pool*sqrt((1/nA)+(1/nB)))

pt(t, df=(nA+nB-2), lower.tail=F)
```

Aplicando t-test obtenemos un valor de pt de un **0.118**, bajo el supuesto de la hipótesis nula, la diferencia de medias es igual a cero, con lo que el valor de t será también igual a cero. Cuanto más se aleje t de ese valor, menos probable será que la diferencia observada se deba al azar. Por lo tanto, podemos concluir que no se puede afirmar que hayan variado los precios respecto de un año al otro.

6. Calcular el p_valor en para la misma prueba del problema anterior, usando el método de permutaciones.

```
mean(c(29.29, 22.70))-mean(c(13.57, 14.80, 22.36))

dif_per = NULL
for (i in 1:1000)
{rnd <- sample(1:5,3)
dif_per[i] = mean(x[rnd])-mean(x[-rnd])}
sum(dif_per>=9.08)/length(dif_per) #0.101
```

En primer lugar, calculamos la diferencia entre las medias de los dos grupos, el cual ha sido divido el original de tamaño 5 en dos de 3 y 2. Seguidamente, se combinan todas las observaciones sin tener en cuenta el grupo al que pertenecen. Para cada permutación se calcula la diferencia entre medias hasta un total de 1000 veces. Una vez finalizado el proceso se obtine el *p-value* como la proporción de permutaciones en las que, el valor absoluto de la diferencia calculada, es mayor o igual al valor absoluto de la diferencia observada. Finalmente, obtenemos un resultado de **0.101**, un valor bastante cercano al del ejercicio 5 (0.118).

7. Sabemos que la probabilidad de compra de un producto en el canal internet es de 0.02.En un mes se han conectado 2300 visitantes, de los cuales 94 han comprado nuestro producto, puedo pensar que ha habido un incremento en la probabilidad de compra por internet?

   Probabilidad de comprar = 0.02

   Visitantes conectados = 2300

   Compras = 94

   La probabilidad de compra en ese mes es de 94/2300, donde obtenemos un valor probabilístico de **0.04**, con esto podemos concluir que ha habido un pequeño incremento de compra por internet.

8. Por otro lado, en el mismo mes de la pregunta anterior se ha lanzado una campaña de márqueting directo con un target preseleccionado de 1000 clientes potenciales, obteniendo una respuesta positiva, esto es la compra del producto, en 56 casos. ¿Podemos afirmar que la tasa de respuesta obtenida en el target preseleccionado es mejor que la obtenida por internet.?

   La probabilidad de compra en este mes utilizando el plan de marketing es de 56/1000, donde obtenemos un valor probabilístico de **0.056**, con esto podemos concluir que ha habido un pequeño incremento de compra por la campaña de marketing.

9. Un día me encuentro con un amigo al que hace tiempo que no veía, va acompañado por un hijo varón. Me dice pero que tiene dos hijos. Cuál es la probabilidad de que su otro hijo sea también varón.

   La probabilidad de que su otro hijo sea varón es de un **50%**, ya que únicamente puede ser o varón o mujer, es decir, ½.

10. Hace mucho tiempo, cuanto la televisión era en blanco y negro, empezó en TVE un programa concurso de gran éxito, se llamaba “Un, dos, tres, responda otra vez”. Su primer presentador fue el gran Kiko Ledgard. Una situación típica en dicho programa era cuando al concursante se le ofrecían tres puertas, detrás de una sola de las cuales había el premio. El concursante escogía una de las puertas, y entonces Kiko Ledgard abría una de las dos puertas no escogidas en donde NO había el premio y preguntaba al concursante si quería cambiar de opción (problema de Monty Hall en honor de su creador). ¿Cuál es la mejor opción para el concursante, mantenerse en su primera opción o cambiar de puerta?.

    La mejor opción para el concursante es **cambiar de puerta** para maximizar la probabilidad de ganar el premio. Ya que en la primera elección la probabilidad de cada puerta es de 1/3 (2/3 de perder) pero cuando el presentador te deja cambiar de puerta, al solo haber dos puertas (ya que el presentador abre una puerta donde NO esta el premio), si mantiene su elección inicial mantiene ⅓ de probabilidades de acierto. Por otra parte, si cambia su elección la probabilidad de ganar aumenta a ⅔.