**Tarea 08 - Análisis de datos**
**Giovanni Gamaliel López Padilla**

# Problema 2

**Para investigar si una moneda es justa o no, alguien decide que va a lanzar la moneda 100 veces. Si el número de veces de obtener sol es entre 42 % y 58 %, va a apoyar la hipótesis de que la moneda es justa. Calcula el nivel de significancia correspondiente.**

Sea X una variable aleatoria tal que $X \sim Bern(0.5)$, entonces se tiene que $\mu = 0.5$ y $\sigma^2 = 0.25$. Entonces el valor de t para la distribución t-student es:

$$Z_{\alpha/2} = \frac{Ci - \mu}{\frac{\sigma}{\sqrt{n}}}$$

entonces

$$\begin{aligned} Z_{\alpha/2} &= \frac{0.58 - 0.5}{\frac{0.5}{\sqrt{100}}} \\ &= \frac{0.08}{0.05} \\ Z_{\alpha/2} &= 1.6 \end{aligned}$$

por lo tanto usando la función pnorm(1.6,lower.tail=F) con $Z_{\alpha/2}$ se obtiene que el nivel de significancia es $\alpha = 0.1095986$.

# Problema 3

**Por experiencia se sabe que el número de accesos ,X, durante una hora a una base de datos sigue una distribución Poisson:**

$$P(X = x) = exp(-\lambda)\frac{\lambda^x}{x!} \qquad \text{para } x = 0, 1, 2, \ldots \text{ y } \lambda > 0$$

**Calcula el estimador de máxima verosimilitud para $\lambda$ en una muestra.**

La función de máximo verosimilitud de la distribución es Poisson es:

$$\mathcal{L} = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^x}{x!}$$

Entonces, la función log-verosimilitud es:

$$
\begin{aligned}
log(\mathcal{L}) &= log\left(\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^x}{x!}\right) \\
&= \sum_{i=1}^{n} log\left(e^{-\lambda}\frac{\lambda^x}{x!}\right) \\
&= \sum_{i=1}^{n} \left[log(e^{-\lambda}) + log(\lambda^{x_i}) - log(x!)\right] \\
&= \sum_{i=1}^{n} \left[-\lambda + x_i log(\lambda) - log(x_i)\right] \\
log(\mathcal{L}) &= -\lambda n + log(\lambda)\sum x_i - \sum log(x_i!)
\end{aligned}
$$

Por lo que calculando el valor crítico de $log(\mathcal{L})$ se obtiene que:

$$
\begin{aligned}
\frac{\partial \log(\mathcal{L})}{\partial \lambda} &= 0 \\
-n + \frac{\sum x_i}{\lambda} &= 0 \\
\hat{\lambda} &= \frac{1}{n}\sum x_i
\end{aligned}
$$

donde $\hat{\lambda}$ corresponde a la media aritmética.

Calculando la segunda derivada de $log(\mathcal{L})$ para comprobar que el valor crítico ($\hat{\lambda}$) se trata de un máximo se obtiene lo siguiente:

$$
\begin{aligned}
\frac{\partial^2 \log(\mathcal{L})}{\partial \lambda^2} &< 0 \\
-\frac{1}{\lambda^2}\sum x_i &< 0
\end{aligned}
$$

como $\lambda^2 > 0$ y $x_i \in \{0, 1, 2, \dots\}$, entonces, la segunda derivada de $log(\mathcal{L})$ será siempre negativa, por lo tanto la media arítmetica es el estimador de máxima verosimilitud.

# Problema 4

**Considera los siguientes datos de un estudio en Bélgica sobre la intención de voto entre 1000 parejas. Las variables $X_1$ , $X_2$ indican si la mujer, respectivamente el hombre, votará para un partido de la coalición (0) o de la oposición (1) en caso de que hubieran elecciones en ese momento.**

| | $X_1 = 0$ | $X_1 = 1$ |
|---|---|---|
| $X_2 = 0$ | 245 | 170 |
| $X_2 = 1$ | 218 | 367 |

**Tabla 1:** Datos recolectados.

**a)** Calcula el oddsratio $\hat{R}$

Con los datos de la tabla 1 se tiene lo siguiente:

$$\begin{aligned}\hat{R} &= \frac{245 * 367}{218 * 170} \\ &= \frac{89915}{37060} \\ \hat{R} &= 2.426201\end{aligned}$$

**b)** Se puede mostrar que si el tamaño de la muestra va a $\infty$, la distribución de $\log(\hat{R})$ converge a una normal con promedio $\log(\hat{R})$, el verdadero log-oddsratio de la distribución subyacente, y con varianza

$$\frac{1}{n_{0,0}} + \frac{1}{n_{0,1}} + \frac{1}{n_{1,0}} + \frac{1}{n_{1,1}}$$

donde $n_{i,j}$ es el número de observaciones con $X_1 = i$ y $X_2 = j$. ¿Apoyas la hipótesis que la pareja vota de manera independiente ($\alpha = 0.05$)?

Calculando los intervalos de confianza, se tiene que:

$$\begin{aligned}\sigma^2 &= \frac{1}{n_{0,0}} + \frac{1}{n_{0,1}} + \frac{1}{n_{1,0}} + \frac{1}{n_{1,1}} \\ &= \frac{1}{245} + \frac{1}{367} + \frac{1}{218} + \frac{1}{170} \\ &= 0.01727594\end{aligned}$$

entonces:

$$\begin{aligned}cu &= exp(log(\hat{R}) + z_{\alpha/2}\sigma) \\ &= exp(0.8863266 + (1.644854)(0.131438)) \\ &= exp(1.102523) \\ cu &= 3.011754 \\ ci &= exp(log(\hat{R}) - z_{\alpha/2}\sigma) \\ &= exp(0.8863266 - (1.644854)(0.131438)) \\ &= exp(0.6701303) \\ ci &= 1.954492\end{aligned}$$

entonces el intervalo de confianza es $[1.954492, 3.011754]$. Como el límite inferior es mayor a 1, entonces no se apoya la hipótesis que la pareja vota de manera independiente.

# Problema 5

**Sea c una cierta cadena binaria de longitud 100. Se quiere verificar si proviene de una muestra $Bern(0.5)(= H_0)$. Para eso se calcula el número de cambios. Un cambio es un 1 seguido por un 0 o un 0 seguido por un 1 en la cadena.**

**a)** Calcula T el número de cambios en una cadena usando una sola linea de código en R.

Para realizar el calculo del número de cambios se hizo en base a la diferencia del i-esimo elemento y el i+1. En el caso en que este sea 00 o 11, la diferencia es 0. Por ende el 0 representa que no existe un cambio. En cambio 01 y 10, produce una diferencia de -1 y 1 respectivamente. Por ende, al tomar el valor absoluto, 1 representa que existio un cambio en la cadena. Al realizar la suma de estas diferencias se obtiene el número de cambios totales en la cadena. Por lo tanto, el código para conseguir el número de cambios en una cadena es el siguiente:

```
change <- sum(abs(diff(chain)))
```

**b)** Usando muchas simulaciones de cadenas bajo $H_0$ , estima y visualiza la distribución de T.

La distribución usando un espacio de 1000 bootstrap se representa en la figura 1.

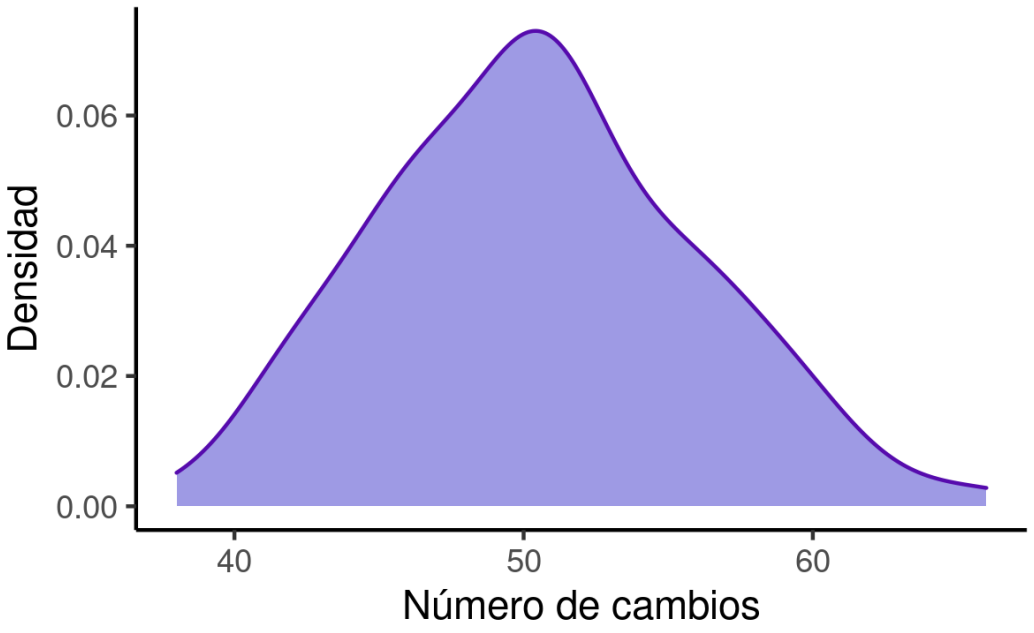



**Figura 1:** Distribución del número de cambios en una cadena de 100 elementos.

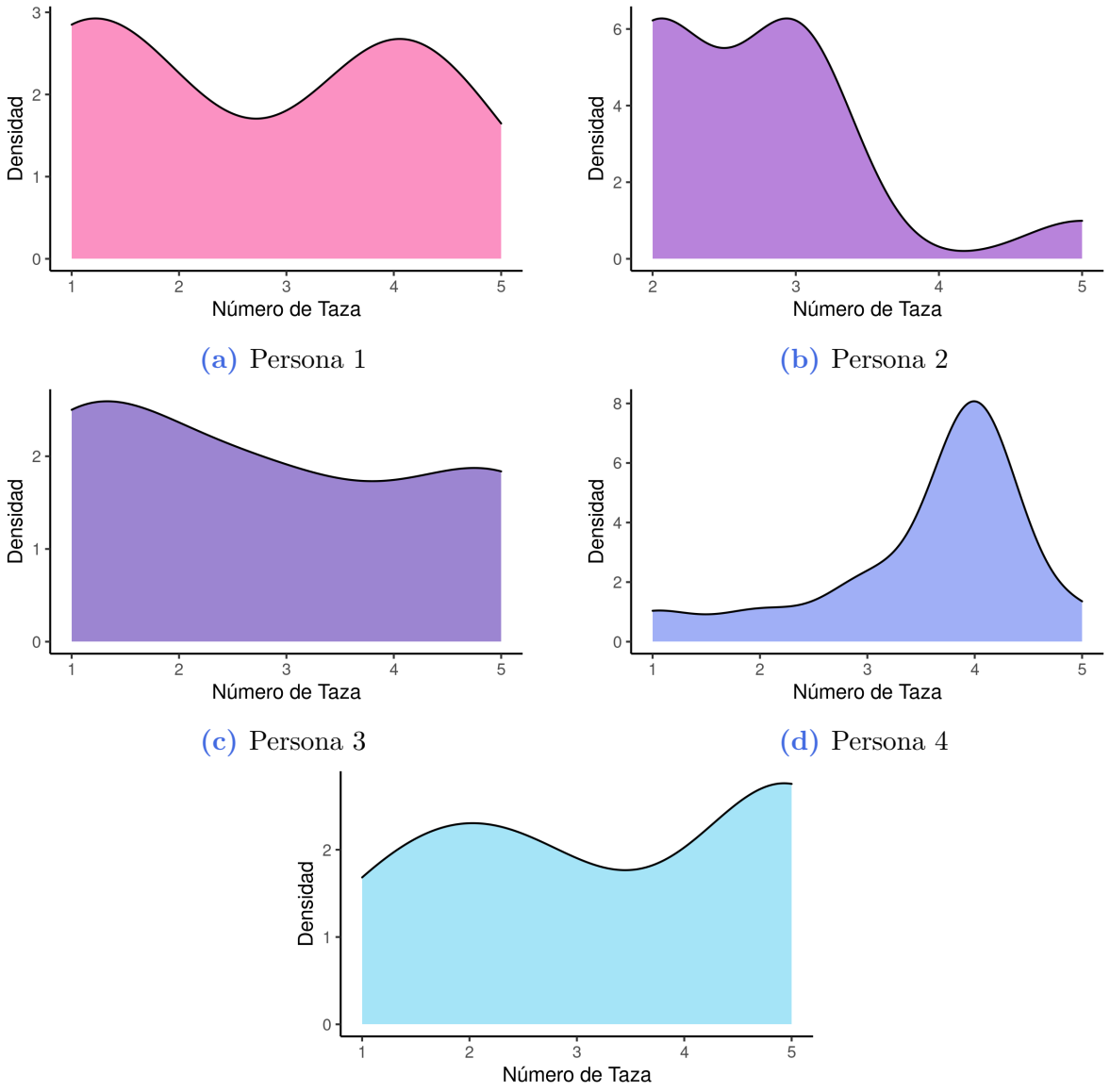
**c)** Calcula el valor de p para $H_0$ si T=42.

Se tiene que la media y varianza de la cadena binaria es 50.5000 y 5.4542 respectivamente. Por lo que calculando el valor de p con el comando pt(mu-42)/(sigma/sqrt(n)), se obtiene que es $9.384e-290$.

# Problema 6

**¿Hay alguna semejanza entre una taza y su dueño? Para eso se decide hacer un pequeño experimento. Se muestran a n voluntarios 5 fotos de personas y 5 tazas en orden al azar. Se pide a cada persona asociar cada taza con una persona (una a una). ¿Cómo formular una prueba de hipótesis para este problema? Propon una estadistica de prueba. Estima su distribución con simulaciones de respuestas bajo $H_0$.**

Una manera de realizar una prueba de hipótesis para este problema es hacer uso del método de bootstrap para obtener una distribución para cada persona en la encuesta. Los valores que podriamos esperar es un máximo en la taza que tenga la probabilidad de que la persona sea dueña de ella. Estos máximos deben ser diferentes en todas las personas, dando así una distribución chicuadrada. Esto debido a que el máximo no necesariamente debe estar en el centro de la distribución.




(a) Persona 1

(b) Persona 2

(c) Persona 3

(d) Persona 4

(e) Persona 5

**Figura 2:** Distribución de la elección de la taza por cada personaa partir de la encuesta realizada.

# Problema 7

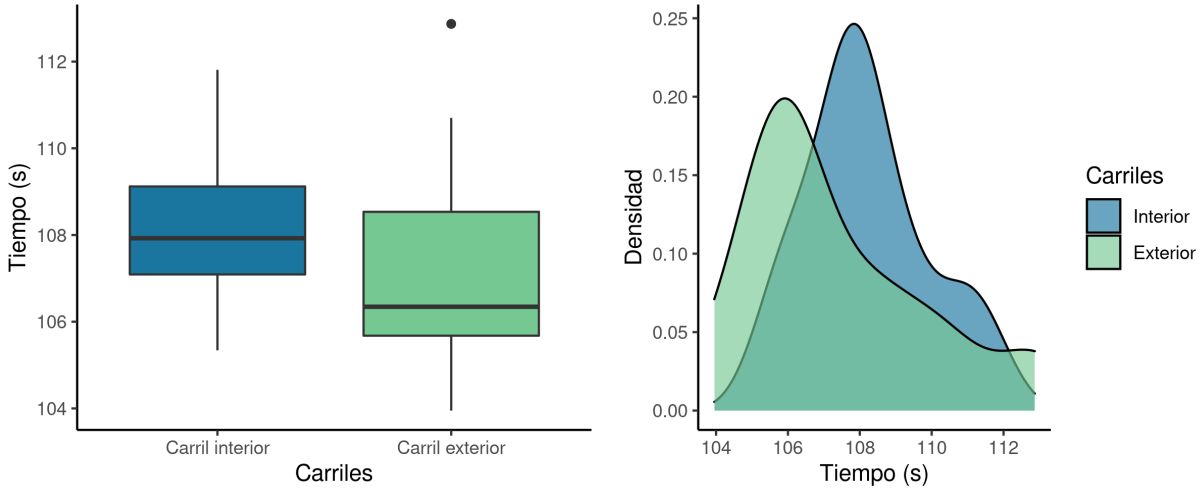
**Durante los juegos olimpicos de Salt Lake City surgió en un periódico la discusión si en las pruebas de 1500m de patinaje, la persona en el carril exterior no tendria ventaja sobre el carril interior. Se organizaron 24 pruebas (una se canceló por una caida). Abajo los tiempos. Aplica una(s) pruebas de estadistica relevante para contestar esta pregunta.**

En la tabla 2 se encuentran las medidas de tendencia central y despersión de los datos contenidos en el archivo data.csv. Se observa que estas medidas son semejantes entre los dos conjuntos, en donde existe una mayor diferencia es en la varianza.

| Linea | Mediana (s) | Promedio (s) | Varianza ($s^2$) |
|---|---|---|---|
| Interna | 107.9 | 108.2 | 2.9653 |
| Externa | 106.3 | 107.4 | 6.3327 |

**Tabla 2:** Medidad de tendencia central y de dispersión de los datos.

Una primera aproximación para llegar a una conclusión es visualizar la distribución de datos por medio de boxplots (figura 3a) y las frecuencias de los datos (figura 3b). Con esto podemos comprobar resultado obtenido en la tabla 2. Con esto podemos llegar a la aproximación que la hipótesis será rechazada. Esto debido a que los datos presentan distribuciones diferentes.




**(a)** Boxplot de los datos del carril interior y exterior.

**(b)** Distribución de los datos del carril interior y exterior.

**Figura 3:** Representación gráfica de la distribucion de los datos.

En la tabla 2 se observa que las varianzas muestrales son diferentes, por ende, al estar basadas en un estimador insesgado, entonces las varianzas de los datos son diferentes. Entonces el nonpooled variance (agregar cita) puede ser calculado como:

$$S_d^2 = \frac{S_x^2}{n} + \frac{S_y^2}{n}$$

El cual es un estimador insesgado para $Var(\bar{X} - \bar{Y})$, entonces, se puede calcular un valor t de la siguiente manera:

$$T_d = \frac{\bar{X} - \bar{Y}}{S_d}$$

El calculo de $T_d$ con los datos, da como resultado 1.2343. El cual es un valor lejano a 0, por ende, la hipótesis es rechaza. Otra manera de comprobar esto es obteniendo el intervalo de confianza para las diferencias de tiempos de una misma carrera. En la tabla 3 se muestran los intervalos de confianza para diferentes valores de $\alpha$. La hipótesis es rechazada unicamente para $\alpha = 0.1$, esto es debido a que el el valor de 0 esta contenido en el intervalo. En cambio para $\alpha = 0.05, 0.01$ la hipótesis no puede ser rechazada.

| $\alpha$ | **Límite inferior** | **Límite superior** |
|---|---|---|
| 0.1 | -1.5177627 | -0.1249646 |
| 0.05 | -1.66300079 | 0.02027351 |
| 0.01 | -1.9672386 | 0.3245113 |

**Tabla 3:** Intervalos de confianza para difentes valores de $\alpha$.

La frecuencia de las diferencias de tiempos se muestra en la figura 4.

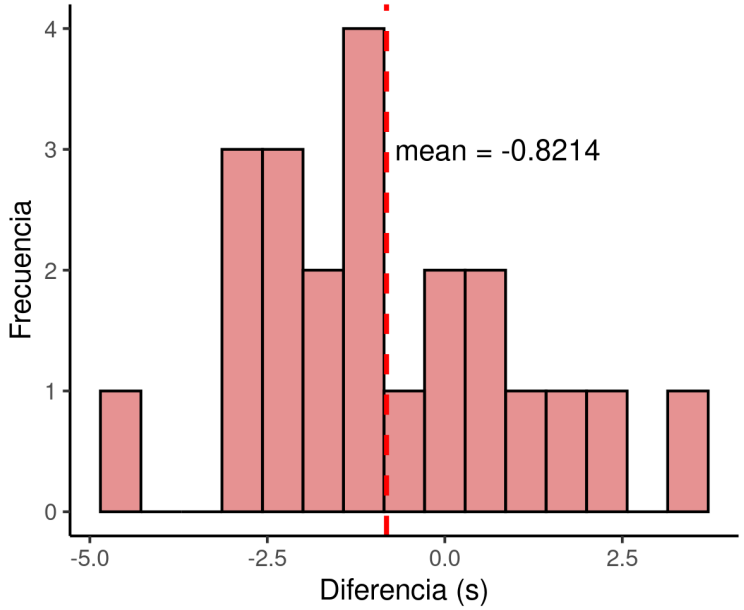



**Figura 4:** Frecuencia de la diferencia entre los tiempos del carril exterior y carril interior.