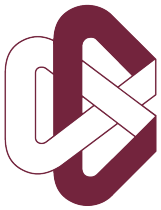


CIMAT

**Centro de Investigación en Matemáticas**
Unidad Monterrey

# Cómputo Estadístico
# Tarea 1

Gustavo Hernández Angeles

27 de agosto de 2025

# Índice

# Problema 1

**1**

La siguiente tabla muestra los resultados parciales de dos encuestas que forman parte de un estudio para evaluar el desempeño del Primer Ministro del Canadá. Se tomó una muestra aleatoria de 1600 ciudadanos canadienses mayores de edad y en los renglones se observa que 944 ciudadanos aprobaban el desempeño del funcionario, mientras que las columnas muestran que, seis meses después de la primera encuesta, sólo 880 aprueban su desempeño.

**Resultados Parciales de Encuestas**

| | **Segunda encuesta** | | |
|---|---|---|---|
| | **Y=1, Aprueba** | **Y=0, Desaprueba** | **Total** |
| **Primera encuesta** | | | |
| **x=1, Aprueba** | 794 | 150 | 944 |
| **x=0, Desaprueba** | 86 | 570 | 656 |
| **Total** | 880 | 720 | 1600 |

a) Considere el modelo de regresión logística

$$\log \frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)} = \beta_0 + \beta_1 x_i$$

Escriba la logverosimilitud correspondiente. Muestre explícitamente (i.e. maximizando la logverosimilitud), que el estimador máximo verosimilitud para $\beta_1$ es el logaritmo de la tasa de momios de la tabla dada (En general, en regresión logística los estimadores de máxima verosimilitud no tienen una forma explícita, sin embargo, en el presente caso si).

b) Sea $p_1$ la proporción de ciudadanos que aprueban el desempeño del ministro al tiempo inicial y sea $p_2$ la proporción correspondiente seis meses después. Considere la hipótesis $H_0 : p_1 = p_2$, ¿Cómo puede hacerse esta prueba?

**Solución:**

**Inciso a)**

Sea $\pi_x = P(Y = 1|x)$. El modelo implica dos ecuaciones, una para cada valor de $x$:

- Si $x = 0$ (desaprobaban en la primera encuesta): $\log \frac{\pi_0}{1-\pi_0} = \beta_0$.
- Si $x = 1$ (aprobaban en la primera encuesta): $\log \frac{\pi_1}{1-\pi_1} = \beta_0 + \beta_1$.

Los datos de la tabla se pueden ver como el resultado de dos muestras binomiales independientes: una de $n_1 = 944$ individuos que aprobaron inicialmente ($x = 1$) y otra de $n_0 = 656$ que desaprobaron ($x = 0$). La variable de respuesta $Y$ es si aprueban en la segunda encuesta.

Denotemos las celdas de la tabla como $n_{xy}$, donde $x$ es el resultado de la primera encuesta y $y$ el de la segunda.

- $n_{11} = 794$ (Aprueba $\rightarrow$ Aprueba)
- $n_{10} = 150$ (Aprueba $\rightarrow$ Desaprueba)
- $n_{01} = 86$ (Desaprueba $\rightarrow$ Aprueba)
- $n_{00} = 570$ (Desaprueba $\rightarrow$ Desaprueba)

La función de log-verosimilitud, agrupando por los valores de $x$:

$$\ell(\beta_0, \beta_1) = \ell(\pi_0, \pi_1) = [n_{01}\log(\pi_0) + n_{00}\log(1-\pi_0)] + [n_{11}\log(\pi_1) + n_{10}\log(1-\pi_1)]$$

Para maximizar $\ell$, podemos maximizar cada parte por separado. El estimador de máxima verosimilitud para $\pi_0$ (la probabilidad de aprobar en la segunda encuesta, dado que se desaprobaba en la primera) es la proporción muestral:

$$\hat{\pi}_0 = \frac{n_{01}}{n_{01} + n_{00}} = \frac{86}{86 + 570} = \frac{86}{656}$$

El EMV para $\pi_1$ (la probabilidad de aprobar en la segunda encuesta, dado que se aprobaba en la primera) es:

$$\hat{\pi}_1 = \frac{n_{11}}{n_{11} + n_{10}} = \frac{794}{794 + 150} = \frac{794}{944}$$

Usando la propiedad de invarianza de los EMV, podemos estimar $\beta_0$ y $\beta_1$ sustituyendo $\hat{\pi}_0$ y $\hat{\pi}_1$ en las ecuaciones del modelo:

$$\hat{\beta}_0 = \log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right)$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right)$$

Despejando $\hat{\beta}_1$ de la segunda ecuación:

$$\hat{\beta}_1 = \log\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right) - \hat{\beta}_0 = \log\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right) - \log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right)$$

$$\hat{\beta}_1 = \log\left(\frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_0/(1-\hat{\pi}_0)}\right)$$

Mostrando así, que el estimador de $\hat{\beta}_1$ es el logaritmo de la tasa de momios de la tabla dada.

**Inciso b)**

La hipótesis a probar es $H_0 : p_1 = p_2$, donde $p_1$ y $p_2$ son las proporciones poblacionales de aprobación en la primera y segunda encuesta, respectivamente. Los datos no provienen de muestras independientes, sino que son mediciones repetidas sobre los mismos 1600 individuos. Por lo tanto, se trata de datos pareados. La prueba adecuada para comparar proporciones en muestras pareadas es la *prueba de McNemar*.

Esta prueba se centra en los individuos que cambiaron de opinión entre las dos encuestas, es decir, las celdas discordantes de la tabla:

- $n_{10} = 150$: Personas que aprobaron en la primera encuesta pero desaprobaron en la segunda.

- $n_{01} = 86$: Personas que desaprobaron en la primera encuesta pero aprobaron en la segunda.

La hipótesis nula de igualdad de proporciones marginales ($H_0 : p_1 = p_2$) es equivalente a la hipótesis de que la probabilidad de cambiar de opinión en una dirección es igual a la probabilidad de cambiar en la dirección opuesta.

El estadístico de prueba de McNemar se calcula como:

$$\chi^2 = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}$$

Bajo la hipótesis nula, este estadístico sigue una distribución Chi-cuadrada ($\chi^2$) con 1 grado de libertad. Calculamos el valor del estadístico $\chi^2$ con los datos de la tabla.

$$\chi^2 = \frac{(150 - 86)^2}{150 + 86} = \frac{64^2}{236} = \frac{4096}{236} \approx 17.36$$

Comparando el valor del estadístico con el valor crítico de una distribución $\chi^2$ con 1 grado de libertad para un nivel de significancia, digamos $\alpha = 0.05$, el valor crítico es 3.841. En este caso, $17.36 > 3.841$, lo que indica que se rechaza la hipótesis nula $H_0$ y concluímos que hay una diferencia estadísticamente significativa entre la proporción de aprobación en la primera encuesta y la proporción de aprobación en la segunda.

## Problema 2

**2**

Suponga $(x_1, y_1), \ldots, (x_n, y_n)$ observaciones independientes de variables aleatorias definidas como sigue:

$$Y_i \sim \text{Bernoulli}(p), \quad i = 1, \ldots, n$$
$$X_i|\{Y_i = 1\} \sim N(\mu_1, \sigma^2)$$
$$X_i|\{Y_i = 0\} \sim N(\mu_0, \sigma^2)$$

Usando el Teorema de Bayes, muestre que $P(Y_i = 1|X_i)$ satisface el modelo de regresión logística, esto es

$$\text{logit}(P(Y_i = 1|X_i)) = \alpha + \beta X_i$$

con $\beta = (\mu_1 - \mu_0)/\sigma^2$.

**Solución:**

Podemos aplicar directamente el Teorema de Bayes sobre la probabilidad $P(Y_i = 1|X_i)$, tomando en cuenta que al ser $X$ una variable continua, su "probabilidad" se calcula con su funcion de densidad de probabilidad $f_X(X_i|Y_i = y)$.

$$P(Y_i = 1|X_i) = \frac{f_X(X_i|Y_i = 1)P(Y_i = 1)}{f_X(X_i|Y_i = 1)P(Y_i = 1) + f_X(X_i|Y_i = 0)P(Y_i = 0)}$$

Determinamos el momio de esta probabilidad para acercarnos a la forma del logit.

$$\begin{aligned}
\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)} &= \frac{\frac{f_X(X_i|Y_i=1)P(Y_i=1)}{f_X(X_i|Y_i=1)P(Y_i=1)+f_X(X_i|Y_i=0)P(Y_i=0)}}{\frac{f_X(X_i|Y_i=0)P(Y_i=0)}{f_X(X_i|Y_i=1)P(Y_i=1)+f_X(X_i|Y_i=0)P(Y_i=0)}} \\
&= \frac{f_X(X_i|Y_i = 1)P(Y_i = 1)}{f_X(X_i|Y_i = 0)P(Y_i = 0)} \\
&= \frac{\exp\left(-(X_i - \mu_1)^2/2\sigma\right)p}{\exp\left(-(X_i - \mu_0)^2/2\sigma\right)(1 - p)} \\
&= \frac{p}{1 - p} \exp\left[(-(X_i - \mu_1)^2 + (X_i - \mu_0)^2)/2\sigma\right] \\
&= \frac{p}{1 - p} \exp\left[((\mu_1 - \mu_2)X_i + \mu_0^2 - \mu_1^2)/2\sigma\right]
\end{aligned}$$

Ahora aplicamos logaritmo natural para obtener el logit de la probabilidad deseada $\text{logit}\left(P(Y_i = 1|X_i)\right)$.

$$\begin{aligned}
\text{logit}\left(P(Y_i = 1|X_i)\right) &= \log\left(\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}\right) \\
&= \log\left(\frac{p}{1 - p}\right) + \frac{\mu_0^2 - \mu_1^2}{2\sigma} + \frac{\mu_1 - \mu_2}{2\sigma} X_i
\end{aligned}$$

Haciendo $\alpha = \log\left(\frac{p}{1-p}\right) + \frac{\mu_0^2 - \mu_1^2}{2\sigma}$ y $\beta = \frac{\mu_1 - \mu_2}{2\sigma}$, finalmente obtenemos:

$$\text{logit}\left(P(Y_i = 1|X_i)\right) = \alpha + \beta X_i$$

## Problema 3

**3**

La siguiente tabla muestra conteos de células $T_4$ por $mm^3$ en muestras de sangre de 20 pacientes (en remisión) con enfermedad de Hodgkin, así como conteos en 20 pacientes en remisión de otras enfermedades. Una cuestión de interés es si existen diferencias en las distribuciones de conteos en ambos grupos.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| H | 396 | 568 | 1212 | 171 | 554 | 1104 | 257 | 435 | 295 | 397 |
| No-H | 375 | 375 | 752 | 208 | 151 | 116 | 736 | 192 | 315 | 1252 |
| H | 288 | 1004 | 431 | 795 | 1621 | 1378 | 902 | 958 | 1283 | 2415 |
| No-H | 675 | 700 | 440 | 771 | 688 | 426 | 410 | 979 | 377 | 503 |

a) Haga una comparación gráfica exploratoria de estos datos.

b) Ajuste un modelo de Poisson apropiado.

c) Usando la normalidad asintótica de los estimadores de máxima verosimilitud, dé un intervalo del 90 % de confianza para la diferencia en medias. ¿Hay evidencia de diferencias en los dos grupos en cuanto a las medias de los conteos?

### Solución:

#### Inciso a)

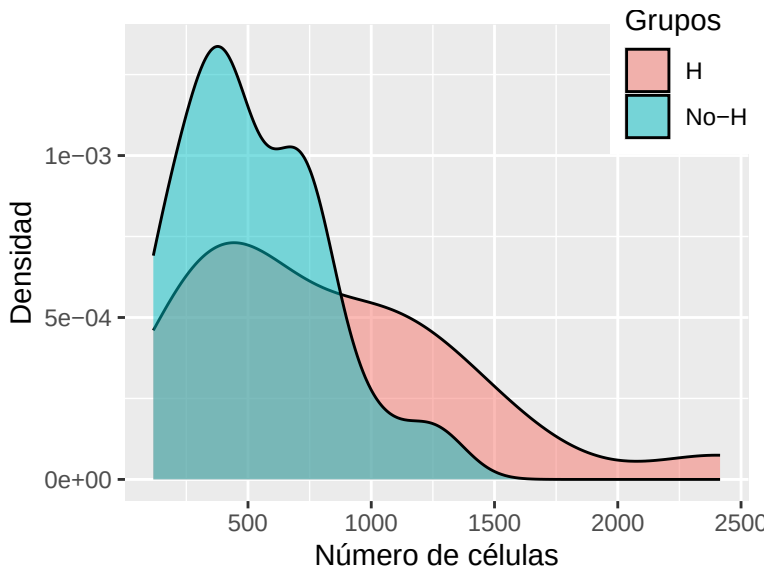
Primero leemos los datos. Estableceremos la variable binaria *has_h* para especificar los pacientes con enfermedad de Hodgkin.

```
library(ggplot2)

h_counts <- c(396, 568, 1212, 171, 554, 1104, 257, 435, 295, 397,
              288, 1004, 431, 795, 1621, 1378, 902, 958, 1283, 2415)
noh_counts <- c(375, 375, 752, 208, 151, 116, 736, 192, 315, 1252,
                675, 700, 440, 771, 688, 426, 410, 979, 377, 503)

df <- data.frame(
group = factor(c(rep("H", length(h_counts)), rep("No-H", length(noh_counts)))),
  counts = c(h_counts, noh_counts)
)
```

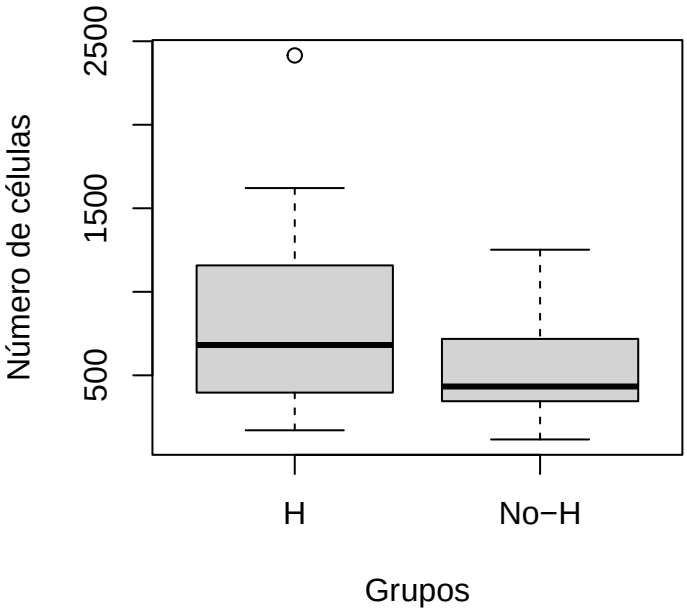
Podemos comparar las distribuciones del conteo de células de los pacientes con y sin la enfermedad de Hodgkin mediante Kernel Density Estimation (KDE) (Figura 3.1) y Boxplot (Figura 3.1).

```
ggplot(data = df) +
  geom_density(aes(x = counts, fill = group), alpha = 0.5) +
  labs(x = "Número de células", y = "Densidad", fill = "Grupos") +
```

```
    theme(legend.position="inside",legend.position.inside = c(0.9, 0.9))
```




```
boxplot(counts ~ group, data=df, ylab="Número de células", xlab="Grupos")
```




El gráfico de densidad muestra una distribución más sesgada "hacia la derecha"para los pacientes con la enfermedad de Hodgkin que se traduce en una mayor varianza y una media muestral mayor respecto al grupo de pacientes sin la enfermedad.

El gráfico de caja muestra los cuartiles y la mediana de las distribuciones. La comparación visual sugiere también una mayor mediana en el conteo de las células $T_4$.

### Inciso b)

En este inciso codificamos la variable binaria para que posea valores 0 y 1, en donde 1 representa a los pacientes con la enfermedad de Hodgkin.

```
df <- data.frame(
  has_h = factor(c(rep(1, length(h_counts)), rep(0, length(noh_counts)))),
  counts = c(h_counts, noh_counts)
)
poisson_model <- glm(counts ~ has_h, data = df, family = poisson)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = counts ~ has_h, family = poisson, data = df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.257763   0.009787   639.4   <2e-16 ***
## has_h1      0.455436   0.012511    36.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 11325  on 39  degrees of freedom
## Residual deviance:  9965  on 38  degrees of freedom
## AIC: 10294
##
## Number of Fisher Scoring iterations: 5
```

El modelo de regresión de Poisson ajustado se especifica de la siguiente manera:

$$\ln(\mathbb{E}[\text{conteo}]) = \beta_0 + \beta_1 \cdot \text{has\_h}$$

Los resultados del modelo, proporcionados en el `summary` de R, nos dan las estimaciones de los coeficientes:

- **Coeficiente del Intercepto** ($\beta_0$): 6.257763
- **Coeficiente para** `has_h1` ($\beta_1$): 0.455436

Estos coeficientes tienen la siguiente interpretación:

- **El Intercepto**: El valor de 6.257763 es el logaritmo del conteo promedio de células $T_4$ para el grupo de pacientes sin la enfermedad de Hodgkin (cuando `has_h = 0`). Para obtener el conteo promedio en la escala original, se calcula $e^{6.257763} \approx 522.2$.
- **El Coeficiente** `has_h1`: El valor de 0.455436 representa el **cambio** en el logaritmo del conteo promedio de células cuando se pasa del grupo sin la enfermedad de Hodgkin (`has_h = 0`) al grupo con la enfermedad (`has_h = 1`). Este valor indica un incremento en el logaritmo de la media.

Para una interpretación más intuitiva, se puede calcular la *Razón de Tasa de Incidencia (RTI)*, la cual se obtiene exponentiando el coeficiente: $e^{0.455436} \approx 1.577$. Esto significa que el conteo promedio de células $T_4$ en los pacientes con la enfermedad de Hodgkin es aproximadamente 1.577 veces mayor que en los pacientes sin la enfermedad.

**Dispersión del Modelo:** Es crucial notar que el modelo presenta sobredispersión. Esto se evidencia al comparar la devianza residual (9965) con los grados de libertad residuales (38). El valor de la devianza es mucho mayor que los grados de libertad, lo que indica que la varianza de los datos es significativamente mayor que la media. Aunque el modelo de Poisson muestra una diferencia significativa, los errores estándar pueden estar subestimados, lo que infla el valor del estadístico $z$ y, por lo tanto, reduce el $p$-valor de manera artificial.

### Inciso c)

Para completar el inciso, se utiliza la propiedad de normalidad asintótica de los estimadores de máxima verosimilitud para construir un intervalo de confianza del 90 % para la diferencia en las medias. En el modelo de Poisson, esta diferencia se modela a través del coeficiente $\beta_1$ que representa el logaritmo de la razón de las medias.

Del resumen de los resultados del modelo de regresión de Poisson, se extraen los siguientes valores para el coeficiente $\beta_1$ (`has_h1`):

- Estimación ($\hat{\beta}_1$): 0.455436
- Error Estándar (SE($\hat{\beta}_1$)): 0.012511

El valor crítico para un intervalo de confianza del 90 % en una distribución normal estándar es $Z_{0.95} = 1.645$. El intervalo de confianza se calcula usando la fórmula de normalidad asintótica:

$$\text{IC}_{90\,\%}(\beta_1) = \hat{\beta}_1 \pm Z_{0.95} \cdot \text{SE}(\hat{\beta}_1)$$

Sustituyendo los valores:

$$\begin{aligned} \text{IC}_{90\,\%}(\beta_1) &= 0.455436 \pm 1.645 \cdot (0.012511) \\ &= 0.455436 \pm 0.02058 \\ &= [0.434856, 0.476016] \end{aligned}$$

Este intervalo corresponde al logaritmo natural de la razón de las medias, $\ln(\lambda_{\text{H}}/\lambda_{\text{No-H}})$. Para obtener el intervalo en la escala original, se aplica la función exponencial a los límites del intervalo anterior:

$$\text{IC}_{90\,\%}(\lambda_{\text{H}}/\lambda_{\text{No-H}}) = [e^{0.434856}, e^{0.476016}] = [1.545, 1.609]$$

El intervalo de confianza del 90 % para la razón de las medias de los conteos de células $T_4$ es $[1.545, 1.609]$. Dado que este intervalo no contiene el valor 1, se concluye que existe evidencia estadística de que las medias de los conteos entre los dos grupos son significativamente diferentes. Específicamente, se estima con un 90 % de confianza que la media de conteo para los pacientes con la enfermedad de Hodgkin es entre 1.545 y 1.609 veces mayor que para los pacientes sin la enfermedad.