# Clase 4
# Estadística: Comparación de Dos Muestras y Regresión Lineal

Pablo Barceló
DCC, Universidad de Chile

## 1. Intervalos de Confianza y Tests para la Diferencia de las Medias de Dos Muestras

**Suposiciones básicas:**

1. $X_1, \ldots, X_m$ es muestra aleatoria tomada desde distribución con valor esperado $\mu_1$ y desviación $\sigma_1$.
2. $Y_1, \ldots, Y_n$ es muestra aleatoria tomada desde distribución con valor esperado $\mu_2$ y desviación $\sigma_2$.
3. Las muestras en $X$ e $Y$ son independientes.

Importante: El valor esperado de $\bar{X} - \bar{Y}$ es $\mu_1 - \mu_2$ y su desviación estándar es $\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$. Podemos utilizar $s_1$ y $s_2$ para estimar $\sigma_1$ y $\sigma_2$, respectivamente.

**Procedimientos de testeo:** Asumimos que ambas distribuciones son normales. Por tanto, tanto $\bar{X}$ como $\bar{Y}$ distribuyen de forma normal. (De otra forma, $\bar{X}$ e $\bar{Y}$ distribuyen aproximadamente normal si hay suficientes muestras). Esto quiere decir que $\bar{X} - \bar{Y}$ también distribuye normal con valor esperado $\mu_1 - \mu_2$ y desviación $\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$. En este caso tenemos entonces:

| | |
|---|---|
| Hipótesis nula: | $H_0 : \mu_1 - \mu_2 = \Delta_0$ |
| Estadística de testeo: | $z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$ |
| Hipótesis alternativa | Región de rechazo para test de nivel $\alpha$ |
| $H_a : \mu_1 - \mu_2 > \Delta_0$ | $z \geq z_\alpha$ |
| $H_a : \mu_1 - \mu_2 < \Delta_0$ | $z \leq -z_\alpha$ |
| $H_a : \mu_1 - \mu_2 \neq \Delta_0$ | $z \leq -z_{\alpha/2}$ o $z > z_{\alpha/2}$ |
| Los valores-$p$ se calculan igual que antes. | |

Ejercicio: Una muestra de tamaño $m = 20$ de un tipo de acero entregó una resistencia promedio $\bar{x} = 29{,}8$. Una segunda muestra de tamaño $n = 25$ de otro tipo de acero entregó una resistencia promedio $\bar{y} = 34{,}7$. Asumiendo que ambas distribuciones son normales con desviación $\sigma_1 = 4$ y $\sigma_2 = 5$, respectivamente, ¿es posible inferir desde estos datos que los valores medios de las resistencias de ambos tipos de acero son distintos? Se pide realizar esto con un nivel de significancia $\alpha = 0{,}01$. Respuesta:

1. Hipótesis nula: $H_0 : \mu_1 - \mu_2 = 0$.
2. Hipótesis alternativa: $H_a : \mu_1 - \mu_2 \neq 0$.
3. Valor de la estadística: $z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{16}{m} + \frac{25}{n}}}$.

4. Región de rechazo: $z \leq -z_{0,005}$ y $z \geq z_{0,005}$, es decir $z \leq -2{,}58$ y $z \geq 2{,}58$.

5. Reemplazo de valores desde la muestra: $z = \frac{\bar{2}9{,}8-\bar{3}4{,}7}{\sqrt{\frac{16}{20}+\frac{25}{25}}} = -3,66$.

6. Decisión: Rechazar $H_0$ ya que $z \leq -2{,}58$.

Notar que en este caso el valor-$p$ es $2(1-\Phi(3{,}66)) \approx 0$, por lo que $H_0$ debería ser rechazada a prácticamente cualquier valor de significancia.

**Errores tipo II:** En este caso podemos utilizar la siguiente tabla:

| Hipótesis alternativa | Error de tipo II $\beta(\Delta')$ para el test de nivel $\alpha$ |
|---|---|
| $H_a : \mu_1 - \mu_2 > \Delta_0$ | $\Phi(\frac{\Delta_0-\Delta'}{\sigma} + z_\alpha)$ |
| $H_a : \mu_1 - \mu_2 > \Delta_0$ | $1 - \Phi(\frac{\Delta_0-\Delta'}{\sigma} - z_\alpha)$ |
| $H_a : \mu_1 - \mu_2 \neq \Delta_0$ | $\Phi(\frac{\Delta_0-\Delta'}{\sigma} + z_{\alpha/2}) - \Phi(\frac{\Delta_0-\Delta'}{\sigma} - z_{\alpha/2})$ |
| Asumimos que $\sigma = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$. | |

Ejercicio: Asuma que la verdadera diferencia entre las medias de las resistencias de los dos tipos de acero es 5 y que queremos que la probabilidad de que el test detecte tal diferencia sea de al menos 90 %. ¿Es capaz el test de realizar esto? Respuesta: La probabilidad de que el test no detecte esta diferencia es el error tipo II cuando $\mu_1 - \mu_2 = 5$ (es decir, la probabilidad de que se quede con $H_0$ dado que $\mu_1 - \mu_2 = 5$). Podemos calcular esta probabilidad como

$$\beta(5) \ = \Phi(\frac{-5}{1{,}34} + 2{,}58) - \Phi(\frac{-5}{1{,}34} - 2{,}58) \ = \ 0{,}1251.$$

Por tanto, la probabilidad de que el test detecte la diferencia en este caso es $1 - \beta(5) = 0{,}8749$. Es decir, se necesitan más muestras para poder asegurarse que esto ocurre con al menos un 90 % de seguridad.

**Distribuciones no normales o varianza desconocida:** En este caso podemos aproximar a $\bar{X} - \bar{Y}$ por una normal si $m, n > 40$. Además, podemos utilizar como estadística de testeo a $z = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}}$.

**Intervalos de confianza:** Un intervalo de confianza al $100(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ corresponde a $\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$. Igualmente podemos reemplazar $\sigma_1$ y $\sigma_2$ por $s_1$ y $s_2$, respectivamente, si $m, n > 40$.

# 2. Correlación de muestras:

Considere una muestra consistente de pares $(x_1, y_1), \ldots, (x_n, y_n)$. Definimos la *covarianza en la muestra* como:

$$S_{xy} \ = \ \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \ = \ \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}.$$

El *coeficiente de correlación en la muestra* se define entonces como:

$$r \ = \ \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{xy}}}.$$

Este coeficiente es un estimador puntual para el coeficiente de correlación $\rho(X, Y)$.

Ejercicio: En algunas partes existe una gran correlación entre las concentraciones de dos elementos contaminantes $x$ e $y$. Los datos se presentan en la siguiente tabla:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| x | 0.066 | 0.088 | 0.12 | 0.05 | 0.162 | 0.186 | 0.057 | 0.100 | 0.112 |
| y | 4.6 | 11.6 | 9.5 | 6.3 | 13.8 | 15.4 | 2.5 | 11.8 | 8.0 |

Calcule el coeficiente de correlación en la muestra.

# 3. Regresión Lineal

Asuma que los parámetros $\theta_1, \ldots, \theta_n$ determinan el crédito $y$ que se le asigna a un cliente y que contamos con una cantidad importante de ejemplos de cuánto crédito se le ha asignado a un cliente basado en estos parámetros. Es decir, tenemos un conjunto de datos $D$ de la forma $\{(\bar{x}_1, y_1), \ldots, (\bar{x}_m, y_m)\}$, donde para cada $1 \leq i \leq m$ se satisface que (a) $\bar{x}_i = (x_i^1, \ldots, x_i^n)$, (b) $x_i^j$ es el valor del parámetro $\theta_j$ para el cliente $i$, y (c) $y_i$ es el valor del crédito otorgado al cliente $i$.

Trataremos de encontrar la "mejor" relación lineal entre los $\theta_j$ y el valor de $y$. Para eso, buscaremos pesos $w_0, w_1, \ldots, w_n$ tal que si $h(\bar{x}_i) = w_0 + w_1 x_i^1 + \ldots w_n x_i^n$ para cada $1 \leq i \leq m$, entonces la desviación (o error) de los $h(\bar{x}_i)$'s con respecto al verdadero valor $y_i$ de la salida es mínimo. Esto nos permitirá predecir valores para el crédito a entregar a un nuevo cliente $\bar{x}$.

Definimos una matriz:

$$X \;=\; \begin{bmatrix} 1 & x_1^1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2^1 & x_2^2 & \ldots & x_2^n \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_m^1 & x_m^2 & \ldots & x_m^n \end{bmatrix}$$

Note que $X$ es de $m \times (n+1)$. Definimos además $Y = [y_1, \ldots, y_m]^T$. Es decir, $Y$ es de $m \times 1$.

Recuerde que queremos encontrar pesos $w_0, \ldots, w_d$ que minimicen el error con respecto a la salida. En el caso del método de mínimos cuadrados este error se define como:

$$\frac{\sum_{i=1}^{m} (h(\bar{x}_i) - y_i)^2}{m}.$$

Es posible demostrar que tal error es equivalente con la siguiente expresión:

$$\frac{1}{m}||XW - Y||^2 \;=\; \frac{1}{m}\big(W^T X^T X W - 2W^T X^T Y + Y^T Y\big),$$

donde $W^T = [w_0, w_1, \ldots, w_n]$. Para encontrar un valor $W$ que minimice esta expresión debemos derivarla e igualarla a 0. La ecuación que obtenemos es:

$$X^T X W - X^T Y \;=\; 0$$

Asumiendo que $X^T X$ tiene inversa $(X^T X)^{-1}$, entonces el valor óptimo es $W = (X^T X)^{-1} X^T Y$.

**Algoritmo de regresión lineal:**

1. Construya las matrices $X$ e $Y$ desde su conjunto de datos.
2. Calcule la inversa $(X^T X)^{-1}$ de $X^T X$.
3. Calcule $W = (X^T X)^{-1} X^T Y$.
4. Retorne $W$.