Estadística II

Tema 3. Comparación de dos poblaciones

Curso 2010/11

Tema 3. Comparación de dos poblaciones

Contenidos

- Comparación de dos poblaciones: ejemplos, datos apareados para la reducción de la variabilidad experimental.
- Muestras Independientes:
  - Comparación de medias, varianzas iguales, poblaciones normales.
  - Comparación de varianzas en poblaciones normales.
  - Sensibilidad de los contrastes anteriores.
  - Comparación de medias, muestras grandes.
  - Comparación de proporciones, muestras grandes.
- Muestras apareadas, comparación de medias, diferencias normales.

Tema 3. Comparación de dos poblaciones

Objetivos de aprendizaje

- Saber distinguir cuándo se está trabajando con muestras independientes o con muestras dependientes apareadas. Conocer en qué situaciones es conveniente trabajar con muestras apareadas.
- Saber plantear el contraste de hipótesis apropiado para corroborar o invalidar la comparación que se quiere analizar.
- Saber construir la regla de decisión oportuna teniendo en cuenta el contraste planteado y el caso en el que nos encontramos (hipótesis asumidas).
- Conocer cuáles son las consecuencias sobre las conclusiones obtenidas de la violación de alguno de los supuestos.

Tema 3. Comparación de dos poblaciones

Referencias en la bibliografía

- Meyer, P. "Probabilidad y aplicaciones estadísticas" (1992)
  - Capítulo ¿?
- Newbold, P. "Estadística para los negocios y la economía" (1997)
  - Capítulo 9 (9.6, 9.7, 9.8)
- Peña, D. "Fundamentos de Estadística" (2001)
  - Capítulo 10 (10.5)

Ejemplos

1. Un investigador quiere saber si una propuesta fiscal es acogida de igual forma por hombres y mujeres.

$$H_0 : p_H = p_M$$
$$H_1 : p_H \neq p_M$$

p_H = proporción de hombres que acogen favorablemente la propuesta

p_M = proporción de mujeres que acogen favorablemente la propuesta

Efecto nivel social, educativo, económico, tendencia política:

aleatorizar

Ejemplos

2. Se quiere hacer un estudio comparativo entre las entidades de crédito federales y estatales de los Estados Unidos en términos del ratio entre el endeudamiento total de la entidad y su activo.

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

$$X = \frac{\text{endeudamiento}}{\text{activo}} \text{ entidades federales}$$
$$Y = \frac{\text{endeudamiento}}{\text{activo}} \text{ entidades estatales}$$

Efecto tamaño y antigüedad: muestras apareadas

Ejemplos

3. Un inversor quiere comparar los riesgos asociados a dos mercados diferentes (A y B), teniendo en cuenta que dicho riesgo se mide por la variabilidad en las fluctuaciones diarias de precios. Para ello se obtienen datos de 21 fluctuaciones diarias para el mercado A y de 16 para el mercado B.

$$H_0 : \sigma_X^2 = \sigma_Y^2$$
$$H_1 : \sigma_X^2 \neq \sigma_Y^2$$

$$X = \text{ fluctuaciones diarias en el mercado A}$$
$$Y = \text{ fluctuaciones diarias en el mercado B}$$

Efecto día: aleatorizar
Efecto situación macroeconómica: mismas condiciones

Ejemplos

4. Antes de lanzar una promoción muy agresiva de un cierto producto dirigida a los hipermercados de grandes superficies, la directora de marketing de la empresa quiere saber si “merece la pena” (si se traduce en un incremento en las ventas del producto en este tipo de establecimientos). Para ello se seleccionan al azar 50 hipermercados de Madrid para llevar a cabo la promoción y recoger datos de ventas antes y después de la promoción.

$$H_0 : \mu_X \geq \mu_Y$$
$$H_1 : \mu_X < \mu_Y$$

$X =$ volumen de ventas en hipermercados antes de la promoción

$Y =$ volumen de ventas en hipermercados después de la promoción

Efecto "llamada": muestras apareadas
Efecto “zona”: aleatorizar

Ejemplos

5. Se quiere comprobar si una promoción publicitaria (campaña B) aumenta el volumen de ventas. Para ello se seleccionan 10 ciudades con comportamientos de consumo similares y en 5 de ellas se sigue con la campaña habitual (campaña A) y en las otras 5 se lanza la campaña B.

$$H_0 : \mu_A \geq \mu_B$$
$$H_1 : \mu_A < \mu_B$$

$$X = \text{ volumen de ventas con la campaña habitual (A)}$$
$$Y = \text{ volumen de ventas con la nueva campaña (B)}$$

Efecto ciudad:

aleatorizar la elección de en qué ciudades se llevaban a cabo cada una de las campañas

Muestras Independientes: Comparación de medias, varianzas iguales, poblaciones normales

Objetivo: Dadas 2 poblaciones normales con la misma variabilidad, pero que pueden diferir en la media, se quiere contrastar la hipótesis de igualdad de medias.

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

- Sean $(X_1, \ldots, X_{n_1})$, $(Y_1, \ldots, Y_{n_2})$ dos m.a.s. de $X \sim N(\mu_X, \sigma^2)$ e $Y \sim N(\mu_Y, \sigma^2)$, respectivamente, independientes entre sí.
- Estimador de la varianza común σ^2:

$$s_P^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}$$

  - Es un estimador insesgado que utiliza toda la información disponible.
  - Pondera las dos estimaciones independientes s_X^2 y s_Y^2 proporcionalmente a su precisión.

- Resultados básicos:
  - $\frac{(n_1-1)s_X^2}{\sigma^2} \sim \chi^2_{n_1-1}$, $\frac{(n_2-1)s_Y^2}{\sigma^2} \sim \chi^2_{n_2-1}$ independientes.
  - Si H_0 es cierta, entonces $\overline{X} - \overline{Y} \sim N(0, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$
- Estadístico del contraste $T(X_1, \ldots, X_{n_1}; Y_1, \ldots, Y_{n_2})$:

$$\frac{\overline{X} - \overline{Y}}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\frac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{(n_1+n_2-2)s_P^2/\sigma^2}{n_1+n_2-2}}} =$$

$$= \frac{Z}{\sqrt{\chi^2_{n_1+n_2-2}/(n_1+n_2-2)}} \sim_{H_0} t_{n_1+n_2-2}$$

- Región crítica

$$R_\alpha = \left\{ (x_1, \ldots, x_{n_1}; y_1, \ldots, y_{n_2}) \,/\, \left| \frac{\overline{X} - \overline{Y}}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \geq t_{n_1+n_2-2;\frac{\alpha}{2}} \right\}$$

- ¿Y si queremos realizar contrastes unilaterales?

$$\begin{array}{l} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{array} \quad R_\alpha = \left\{ (x_1, \ldots, x_{n_1}; y_1, \ldots, y_{n_2}) \,/\, \frac{\overline{X} - \overline{Y}}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{n_1+n_2-2;\alpha} \right\}$$

$$\begin{array}{l} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{array} \quad R_\alpha = \left\{ (x_1, \ldots, x_{n_1}; y_1, \ldots, y_{n_2}) \,/\, \frac{\overline{X} - \overline{Y}}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{n_1+n_2-2;\alpha} \right\}$$

- ¿Y si queremos contrastar en general

$H_0 : \mu_X - \mu_Y = d_0$ $H_1 : \mu_X - \mu_Y \neq d_0$	$H_0 : \mu_X - \mu_Y \leq d_0$ $H_1 : \mu_X - \mu_Y > d_0$	$H_0 : \mu_X - \mu_Y \geq d_0$ $H_1 : \mu_X - \mu_Y < d_0$
$T(X_1, \ldots, X_{n_1}; Y_1, \ldots, Y_{n_2}) = \frac{\overline{X} - \overline{Y} - d_0}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim_{H_0} t_{n_1+n_2-2}$		

con $d_0 \geq 0$?

Ejemplo 5

- Supongamos que $X \sim N(\mu_A, \sigma^2)$, $Y \sim N(\mu_B, \sigma^2)$.
- Se toman dos m.a.s., obteniéndose las siguientes cifras de ventas:

CAMPAÑA A	16	14	42	38	23
CAMPAÑA B	61	33	37	63	65

- Estadístico del contraste: $T = \frac{\overline{X} - \overline{Y}}{s_P \sqrt{\frac{2}{5}}}$.

$$\overline{x} = 26{,}6 \qquad \overline{y} = 51{,}8$$

$$s_X^2 = \frac{\sum_{i=1}^5 x_i^2 - 5\overline{x}^2}{4} = 162{,}8 \qquad s_Y^2 = \frac{\sum_{i=1}^5 y_i^2 - 5\overline{y}^2}{4} = 239{,}2$$

$$s_P^2 = \frac{4s_X^2 + 4s_Y^2}{8} = 201$$

$$t = \frac{26{,}6 - 51{,}8}{\sqrt{(201 \cdot 2)/5}} = -2{,}81$$

Ejemplo 5 (cont.)

- Con un nivel de significación α, rechazaremos $H_0 : \mu_A \geq \mu_B$ si $t = \frac{\overline{x}-\overline{y}}{s_P\sqrt{\frac{2}{5}}} = -2{,}81 < -t_{8;\alpha}$

$$t_{8;0,01} = 2{,}896 \quad t_{8;0,05} = 1{,}860 \quad t_{8;0,1} = 1{,}397$$

  Se rechaza H_0 a los niveles $\alpha = 0{,}1; 0{,}05$, y no se rechaza para $\alpha = 0{,}01$.

- El p-valor del contraste es:

$$p = Pr\{t_8 \leq -2{,}81\} = Pr\{t_8 \geq 2{,}81\} \in (0{,}01; 0{,}025)$$

Muestras Independientes: Comparación de varianzas, poblaciones normales

Objetivo: Dadas 2 poblaciones normales, se quiere contrastar la hipótesis de igualdad de varianzas.

$$H_0 : \sigma_X^2 = \sigma_Y^2$$
$$H_1 : \sigma_X^2 \neq \sigma_Y^2$$

- Sean $(X_1, \ldots, X_{n_1})$, $(Y_1, \ldots, Y_{n_2})$ dos m.a.s. de $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$, respectivamente, independientes entre sí.
- Resultado básico: $\frac{(n_1-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n_1-1}$, $\frac{(n_2-1)s_Y^2}{\sigma_Y^2} \sim \chi^2_{n_2-1}$ indep.

$$\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \sim F_{(n_1-1, n_2-1)}$$

- Estadístico del contraste: Si H_0 es cierta:

$$T(X_1, \ldots, X_{n_1}; Y_1, \ldots, Y_{n_2}) = \frac{s_X^2}{s_Y^2} \sim_{H_0} F_{(n_1-1, n_2-1)}$$

Muestras Independientes: Comparación de varianzas, poblaciones normales

- Región crítica

$$R_\alpha = \left\{ (x_1, \ldots, x_{n_1}; y_1, \ldots, y_{n_2}) \,/\, \frac{s_X^2}{s_Y^2} \leq F_{(n_1-1, n_2-1); 1-\frac{\alpha}{2}} \text{ ó } \frac{s_X^2}{s_Y^2} \geq F_{(n_1-1, n_2-1); \frac{\alpha}{2}} \right\}$$

- Contrastes unilaterales:

$$H_1 : \sigma_X^2 > \sigma_Y^2 \Rightarrow \; R_\alpha = \{ \frac{s_X^2}{s_Y^2} \geq F_{(n_1-1, n_2-1); \alpha} \}$$

$$H_1 : \sigma_X^2 < \sigma_Y^2 \Rightarrow \; R_\alpha = \{ \frac{s_X^2}{s_Y^2} \leq F_{(n_1-1, n_2-1); 1-\alpha} \}$$

Muestras Independientes: Comparación de varianzas, poblaciones normales

Ejemplo 3

Para comparar los riesgos de los mercados A y B se obtienen m.a.s. de 21 cambios de precios diarios para el mercado A y de 16 para el mercado B. Se obtiene:

Mercado A	Mercado B
$\overline{x}_A = 0{,}3$	$\overline{x}_B = 0{,}4$
$s_A = 0{,}25$	$s_B = 0{,}45$

- Estadístico del contraste: $T = \frac{s_A^2}{s_B^2} \sim_{H_0} F_{(20,15)}$
- Se ha obtenido $t = \left(\frac{0{,}25}{0{,}45}\right)^2 = 0{,}309$
- Región de rechazo:

$$R_\alpha = \{t \leq F_{(20,15);1-\frac{\alpha}{2}} \text{ ó } t \geq F_{(20,15);\frac{\alpha}{2}}\}$$

Sólo tenemos tablas de 1 cola al 5 % y al 1 %, ¿Qué hacemos?

Muestras Independientes: Comparación de varianzas, poblaciones normales

Ejemplo 3 (cont.)

- Si tenemos un ordenador: paquete de estadística, o Excel, para obtener los valores críticos, o para calcular el p-valor:

$$p = \text{mín}\Big(2Pr\{T \leq 0{,}309 \mid H_0\}, 2Pr\{T \geq 0{,}309 \mid H_0\}\Big) =$$
$$= 2F_{(20,15)}(0{,}309) = 2 \cdot 0{,}0077677 = 0{,}01553$$

  ¿Para qué niveles de significación no se rechaza H_0?

- ¿Y si no tenemos ordenador?
  Hacer el contraste unilateral con $H_1 : \sigma_1^2 > \sigma_2^2$ poniendo siempre la estimación que haya dado mayor en el numerador. En este caso, $s_B > s_A \Rightarrow$

$$H_0 : \sigma_B^2 \leq \sigma_A^2$$
$$H_1 : \sigma_B^2 > \sigma_A^2$$

  Ahora $t = \frac{1}{0{,}309} = 3{,}236$, y podemos usar las tablas para buscar $F_{(15,20);0{,}05} = 2{,}20$, $F_{(15,20);0{,}01} = 3{,}09$ ¿Qué se concluye?

Muestras Independientes: Sensibilidad de los contrastes

Objetivo: ¿Consecuencias sobre las conclusiones obtenidas del no cumplimiento de las hipótesis de trabajo?

- No Normalidad
  - Comparación de medias: por el TCL las medias tienen siempre una distribución próxima a la normal. OJO!!! valores atípicos.
  - Comparación de varianzas: muy sensible.
- Heterocedasticidad
  - Error tipo I (α): poco sensible si tamaños muestrales similares. Muy sensible para tamaños dispares (más del doble)
  - Error tipo II (β): muy sensible (aumenta la probabilidad de no detectar diferencias)
- No muestra aleatoria: Muy sensible

  Principio de aleatorización: Previene de sesgos sistemáticos en la asignación de unidades muestrales. Para evitar detectar diferencias debidas a otros factores.

Muestras Independientes: Comparación de medias, muestras grandes

Objetivo: Dadas 2 poblaciones cualesquiera, queremos contrastar la hipótesis de igualdad de medias

$$\begin{array}{l} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{array}$$

- Sean $(X_1, \ldots, X_{n_1})$, $(Y_1, \ldots, Y_{n_2})$ dos m.a.s. de X e Y, respectivamente, independientes entre sí, con n_1 y n_2 suficientemente grandes.
- Resultado básico: Método aproximado (TCL)

$$T(X_1, \ldots, X_{n_1}; Y_1, \ldots, Y_{n_2}) = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} \sim_{H_0} N(0,1)$$

Muestras Independientes: Comparación de medias, muestras grandes

- En general, para $d_0 \geq 0$:

$T(X_1,\ldots,X_{n_1};Y_1,\ldots,Y_{n_2}) = \dfrac{\overline{X}-\overline{Y}-d_0}{\sqrt{\frac{s_X^2}{n_1}+\frac{s_Y^2}{n_2}}} \sim_{H_0} N(0,1)$		
$H_1 : \mu_X - \mu_Y \neq d_0$	$H_1 : \mu_X - \mu_Y > d_0$	$H_1 : \mu_X - \mu_Y < d_0$
$R_\alpha = \left\{ \lvert T \rvert \geq z_{\frac{\alpha}{2}} \right\}$	$R_\alpha = \{ T \geq z_\alpha \}$	$R_\alpha = \{ T \leq -z_\alpha \}$

Muestras independientes: comparación de proporciones, muestras grandes

Objetivo: Dadas 2 poblaciones, se quiere contrastar la hipótesis de que la proporción de elementos con un atributo es idéntica en ambas poblaciones.

$$H_0 : p_X = p_Y = p_0$$
$$H_1 : p_X \neq p_Y$$

- Sean $(X_1, \ldots, X_{n_1})$, $(Y_1, \ldots, Y_{n_2})$ dos m.a.s. de ambas poblaciones independientes entre sí, con r_X, r_Y = número de observaciones con dicho atributo obtenidas en cada muestra.

$$\text{Proporciones muestrales: } \hat{p}_X = \frac{r_X}{n_1}, \quad \hat{p}_Y = \frac{r_Y}{n_2}$$

Muestras independientes: comparación de proporciones, muestras grandes

Si H_0 es cierta:

- La mejor estimación de la proporción común p_0 es:

$$\hat{p}_0 = \frac{r_X + r_Y}{n_1 + n_2}$$

- $\hat{p}_X - \hat{p}_Y$ v.a. con $E(\hat{p}_X - \hat{p}_Y) = 0$ y $V(\hat{p}_X - \hat{p}_Y) = V(\hat{p}_X) + V(\hat{p}_Y)$, que estimamos por:

$$\hat{V}(\hat{p}_X - \hat{p}_Y) = \frac{\hat{p}_0(1-\hat{p}_0)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_2}$$

- Si n_1 y n_2 suficientemente grandes $\Rightarrow$ TCL

$$\frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}_0(1-\hat{p}_0)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim_{H_0} N(0,1)$$

Muestras independientes: comparación de proporciones, muestras grandes

En general:

$T(X_1,\ldots,X_{n_1};Y_1,\ldots,Y_{n_2}) = \dfrac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}_0(1-\hat{p}_0)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$		
$H_1 : p_X \neq p_Y$	$H_1 : p_X > p_Y$	$H_1 : p_X < p_Y$
$R_\alpha = \left\{\lvert T\rvert \geq z_{\frac{\alpha}{2}}\right\}$	$R_\alpha = \{T \geq z_\alpha\}$	$R_\alpha = \{T \leq -z_\alpha\}$

Muestras independientes: comparación de proporciones, muestras grandes

Ejemplo 1

- Supongamos que $X \sim Ber(p_H)$, $Y \sim Ber(p_M)$. Se quería contrastar:

$$\begin{array}{l} H_0 : p_H = p_M \\ H_1 : p_H \neq p_M \end{array}$$

- Una m.a.s de 800 hombres reveló que 320 de éstos acogían favorablemente la propuesta, y una m.a.s. de 500 mujeres, que 150 de éstas lo hacían.
- Estadístico del contraste: $T = \frac{\hat{p}_H - \hat{p}_M}{\sqrt{\hat{p}_0(1-\hat{p}_0)}\sqrt{\frac{1}{800}+\frac{1}{500}}}$.

$$\hat{p}_H = \frac{320}{800} = 0{,}4, \quad \hat{p}_M = \frac{150}{500} = 0{,}3$$

$$\hat{p}_0 = \frac{320+150}{800+500} = 0{,}3615$$

Muestras independientes: comparación de proporciones, muestras grandes

Ejemplo 1 (cont.)

- $$t = \frac{0{,}4 - 0{,}3}{\sqrt{0{,}3615(1 - 0{,}3615)}\sqrt{\frac{1}{800} + \frac{1}{500}}} = \frac{0{,}1}{0,02738} = 3{,}65$$
- $z_{0{,}005} = 2{,}57 \Rightarrow$ rechazamos H_0 a un nivel $\alpha = 0{,}01$.
- ¿Qué haremos para $\alpha = 0{,}05; 0{,}1$?
- ¿Qué puedes decir del p-valor del contraste?
- Si construimos un IC al 95 % para $p_H - p_M$, ¿Contendrá al 0?

Muestras apareadas, comparación de medias, diferencias normales

Ejemplo 4

Antes de lanzar una promoción muy agresiva de un cierto producto dirigida a los hipermercados de grandes superficies, la directora de marketing de la empresa quiere saber si “merece la pena”. Para ello se seleccionan al azar 50 hipermercados de Madrid para llevar a cabo la promoción y recoger datos de ventas antes y después de la promoción.

Datos apareados

Provienen de la medición de una variable en el mismo individuo antes y después de la aplicación de un tratamiento.

Muestras apareadas, comparación de medias, diferencias normales

Objetivo

Disponer de medidas por pares tomadas en condiciones muy semejantes para que, a priori, las 2 unidades experimentales que comparamos sean lo más iguales posible.

¿Por qué?

- Reducir la variabilidad poblacional: para detectar diferencias
- Controlar el efecto de otros factores: para evitar achacar diferencias debidas a otros factores (¿otra forma?)

Muestras apareadas, comparación de medias, diferencias normales

Ejemplo 2

Se quiere hacer un estudio comparativo entre las entidades de crédito estatales y federales de los Estados Unidos en términos del ratio entre el endeudamiento total de la entidad y su activo.

Objetivo

Queremos controlar el efecto de otros factores: tamaño y antigüedad. Disponer de medidas por pares tomadas en condiciones muy semejantes para que, a priori, las 2 unidades experimentales que comparamos sean lo más iguales posible.

Muestras dependientes apareadas

Se eligieron 145 parejas de entidades de crédito. Cada pareja contenía una unidad estatal y una federal. Los emparejamientos se hicieron de forma que los 2 miembros fuesen lo más parecidos posible en tamaño y antigüedad

Muestras apareadas, comparación de medias, diferencias normales

¿Más opciones?

Incorporar la información sobre el tamaño y la antigüedad en el análisis

ANÁLISIS DE LA VARIANZA

También permite extender a $k > 2$ poblaciones el contraste de igualdad de medias en poblaciones normales con varianzas iguales.

Muestras apareadas, comparación de medias, diferencias normales

Objetivo: Dadas 2 poblaciones se quiere contrastar la hipótesis de igualdad de medias.

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

- Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una m.a.s. de una población normal bivariante con parámetros μ_X, μ_Y,σ_X^2, σ_Y^2 y ρ.
  Es suficiente con $D_i = X_i - Y_i$, $i = 1, \ldots, n$ m.a.s. de una población normal.
- Si H_0 es cierta, entonces $\overline{D}$ es normal con $E(\overline{D}) = 0$ y $V(\overline{D}) = \frac{\sigma_X^2+\sigma_Y^2-2\sigma_X\sigma_Y\rho}{n}$.
- Estadístico del contraste

$$T(D_1, \ldots, D_n) = \frac{\overline{D}}{s_D/\sqrt{n}} \sim_{H_0} t_{n-1}$$

donde $s_D^2 = \hat{V}(D)$ es la cuasivarianza muestral de las diferencias:

$$s_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1} = \frac{\sum_{i=1}^{n} D_i^2 - n\overline{D}^2}{n-1}$$

Muestras apareadas, comparación de medias, diferencias normales

En general:

$T(D_1, \ldots, D_n) = \dfrac{\overline{D} - d_0}{s_D/\sqrt{n}}$		
$H_1 : \mu_X - \mu_Y \neq d_0$	$H_1 : \mu_X - \mu_Y > d_0$	$H_1 : \mu_X - \mu_Y < d_0$
$R_\alpha = \left\{ \lvert T \rvert \geq t_{n-1;\frac{\alpha}{2}} \right\}$	$R_\alpha = \{ T \geq t_{n-1;\alpha} \}$	$R_\alpha = \{ T \leq -t_{n-1;\alpha} \}$

Muestras apareadas, comparación de medias, diferencias normales

Ejemplo 2

- Para la muestra descrita anteriormente:
  Se eligieron 145 parejas de entidades de crédito. Cada pareja contenía una unidad estatal y una federal. Los emparejamientos se hicieron de forma que los 2 miembros fuesen lo más parecidos posible en tamaño y antigüedad
  Se obtuvieron unas diferencias (federal menos estatal) medias de 0,0518, con una desviación típica de 0,3055.
- Estadístico del contraste: $t = \frac{0{,}0518}{0{,}3055/\sqrt{145}} = 2{,}0417$
- $n - 1$ es muy grande, podemos trabajar con los valores críticos de la normal y aproximar el p-valor del contraste por:

$$p - valor = 2P\{Z \geq 2{,}04\} = 2 \cdot 0,0207 = 0,0414$$

Comparación de dos poblaciones

Resumen para dos m.a.s. independientes, contrastes bilaterales

Diferencia de	Hipótesis	Estadístico	Región Rechazo
Medias	Datos normales Var. iguales	$\frac{\overline{X}-\overline{Y}}{s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim_{H_0} t_{n_1+n_2-2}$	$\{\lvert T\rvert \geq t_{n_1+n_2-2;\frac{\alpha}{2}}\}$
	D. no normales Muestras grandes	$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{s_X^2}{n_1}+\frac{s_Y^2}{n_2}}} \sim_{H_0} N(0,1)$	$\{\lvert T\rvert \geq z_{\frac{\alpha}{2}}\}$
Proporciones	Muestras grandes	$\frac{\hat{p}_X-\hat{p}_Y}{\sqrt{\hat{p}_0(1-\hat{p}_0)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim_{H_0} N(0,1)$	$\{\lvert T\rvert \geq z_{\frac{\alpha}{2}}\}$
Varianzas	Datos normales	$\frac{s_X^2}{s_Y^2} \sim_{H_0} F_{(n_1-1,n_2-1)}$	$\{T \leq F_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}$ ó $T \geq F_{(n_1-1,n_2-1);\frac{\alpha}{2}}\}$

$$s_P^2 = \frac{(n_1-1)s_X^2+(n_2-1)s_Y^2}{n_1+n_2-2}$$