Curso Básico de Estadística con R

Prof. Vanesa Jordá

Departamento de Economía
Facultad de Ciencias Económicas y Empresariales
Universidad de Cantabria

Índice

Introducción

La inferencia estadística permite establecer conclusiones sobre una población a partir de una muestra de datos de la misma.

Conceptos clave:

- Parámetro: valor poblacional desconocido sobre el que queremos establecer conclusiones a partir de una muestra.
- Estimador: función de la muestra que se emplea para estimar el parámetro.
- Estimación: valor del estimador para una muestra determinada.

Toda estimación puntual debe de ir acompañada de una medida del error o bien de una horquilla de valores entre los que se encuentre el verdadero valor del parámetro con un determinado nivel de confianza.

Inferencia para la diferencia de medias

Vamos a analizar si existe una brecha salarial entre hombres y mujeres en España. Para ello empleamos datos sobre los salarios de ambos sexos que provienen de la Encuesta de Condiciones de Vida en el año 2013 y que encontraréis en el archivo "Datos.xlsx"

Cuestiones

1. ¿Hay diferencia entre el salario de hombres y mujeres? Sí, en media los hombres encuestados ganan 3254.91 euros más que las mujeres.
2. ¿Es menor el salario en el caso de las mujeres?

Inferencia para la diferencia de medias

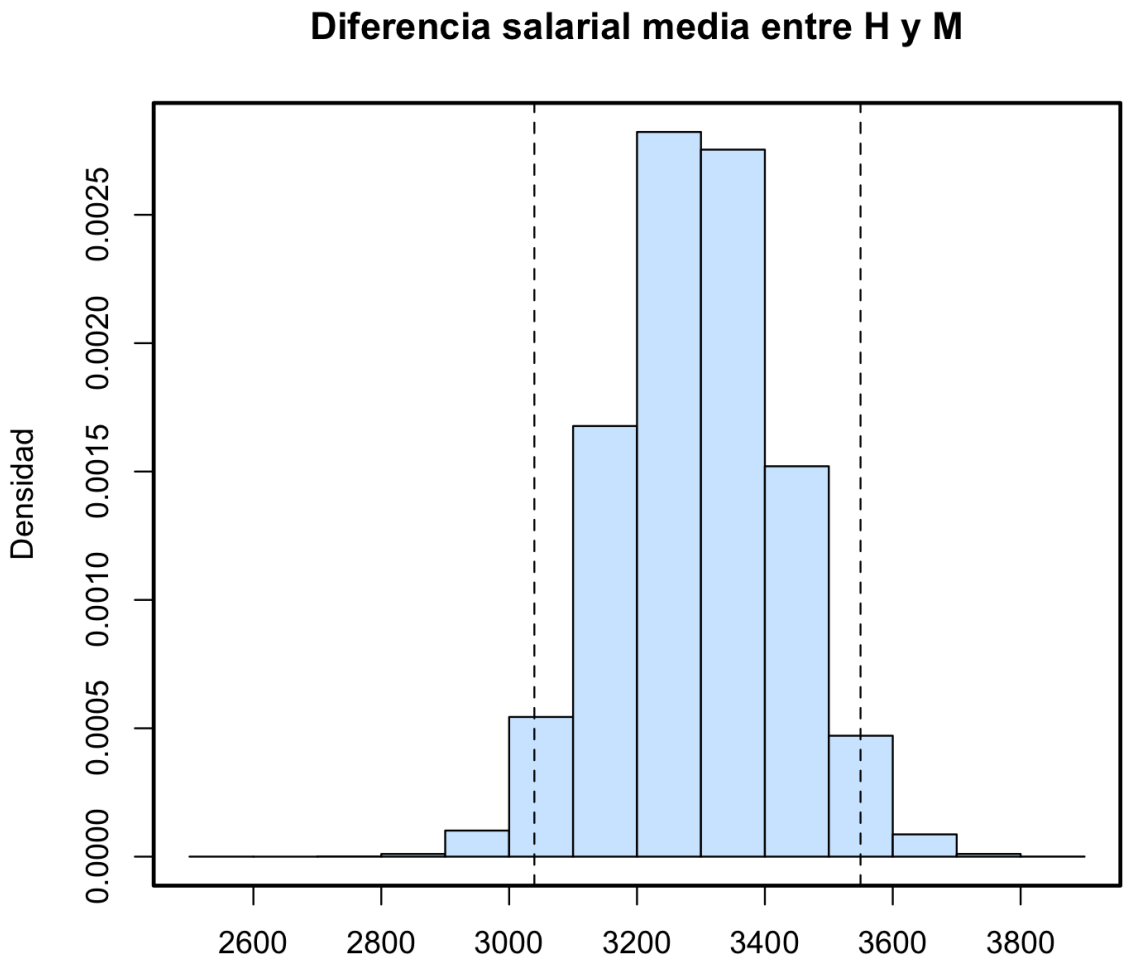
Vamos a analizar si existe una brecha salarial entre hombres y mujeres en España. Para ello empleamos datos sobre los salarios de ambos sexos que provienen de la Encuesta de Condiciones de Vida en el año 2013 y que encontraréis en el archivo "Datos.xlsx"

Cuestiones

1. ¿Hay diferencia entre el salario de hombres y mujeres? Sí, en media los hombres encuestados ganan 3254.91 euros más que las mujeres.
2. ¿Es menor el salario en el caso de las mujeres?

Inferencia para la diferencia de medias

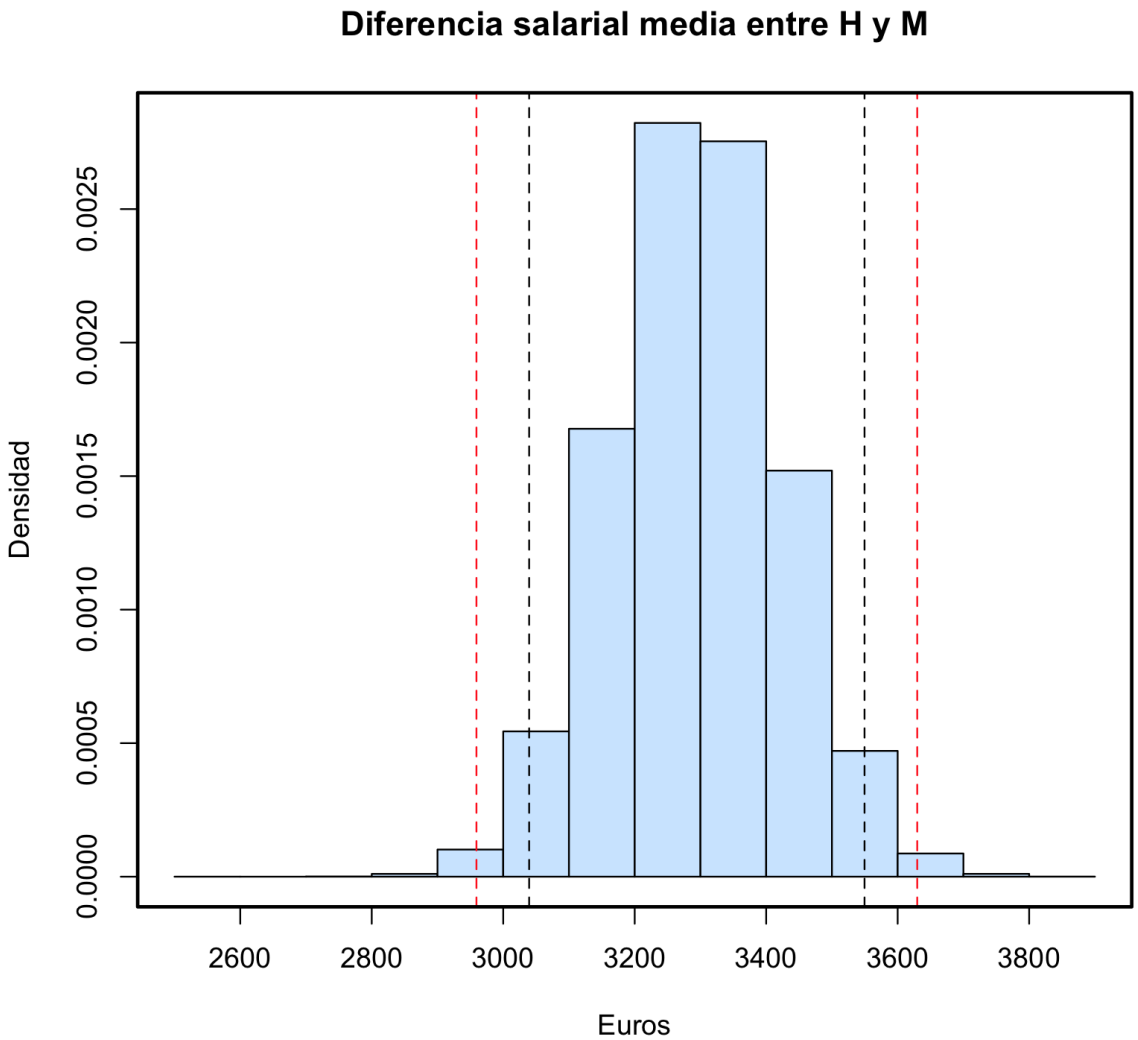
¿Y si tuviésemos 10^6 muestras?



Intervalos de confianza

- Los intervalos de confianza nos proporcionan un rango de valores entre los que el parámetro poblacional se encontrará con un determinado nivel de confianza.
- El nivel de confianza, que denotaremos como $1-\alpha$, suele fijarse en el 95 por ciento, aunque puede variar según las necesidades del análisis.
- Una interpretación alternativa y quizás más intuitiva, es que el intervalo de confianza al 95 por ciento incluye un 95 por ciento de las diferencias de medias muestrales entre hombres y mujeres.
- Otros niveles de confianza que suelen emplearse comúnmente son el 90 por ciento y el 99 por ciento
- ¿Por qué no calcular un intervalo de confianza del 100 por cien?

Inferencia para la diferencia de medias



Elementos de los contrastes de hipótesis

- Especificación del contraste
  1. Hipótesis nula: se asume que es la correcta e incluye los símbolos $=, \leq, \geq$
  2. Hipótesis alternativa: es la que puede ser o no evidenciada por los datos

$$H_0 : \theta = \theta_0$$
$$H_a : \theta \neq \theta_0$$

$$H_0 : \theta \geq \theta_0$$
$$H_a : \theta < \theta_0$$

$$H_0 : \theta \leq \theta_0$$
$$H_a : \theta > \theta_0$$

Contrastes de hipótesis sobre la diferencia de medias

Vamos a analizar si existe una brecha salarial entre hombres y mujeres en España. Por lo que queremos responder a la pregunta...
¿Hay diferencia entre el salario de hombres y mujeres?

Especificación del contraste

$$H_0 : \mu_M = \mu_H \Rightarrow \mu_H - \mu_M = 0$$
$$H_a : \mu_M \neq \mu_H \Rightarrow \mu_H - \mu_M \neq 0$$

Estadístico de contraste:

$$t = \frac{\bar{x}_H - \bar{x}_M}{\sqrt{\frac{S_H^2}{n_H} + \frac{S_M^2}{n_M}}}$$

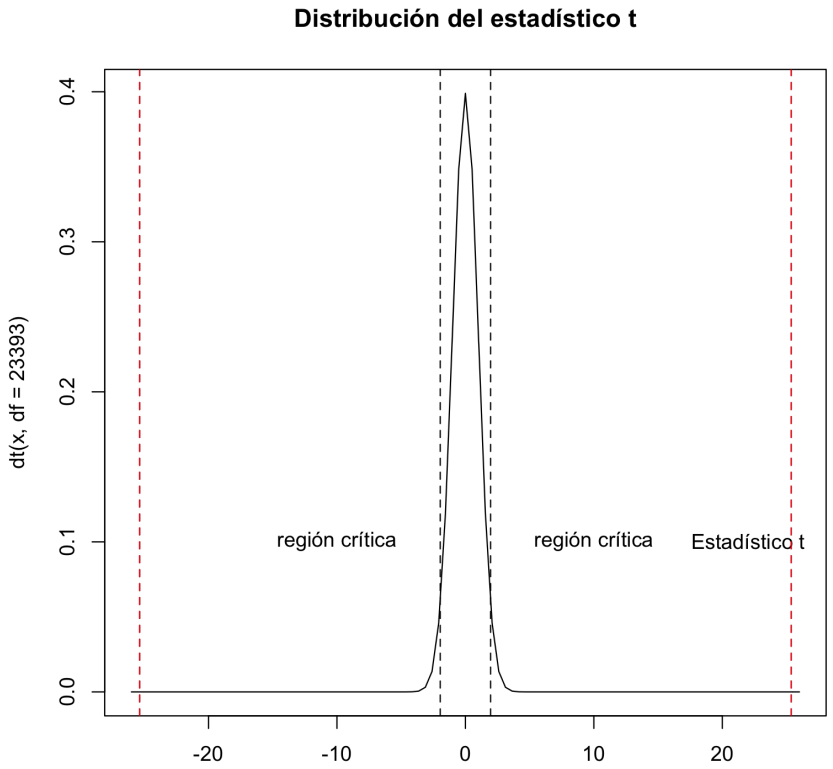
Regla de decisión:

$$|t| > t_{n_H+n_M-2,1-\alpha/2} \Rightarrow RechazoH0$$

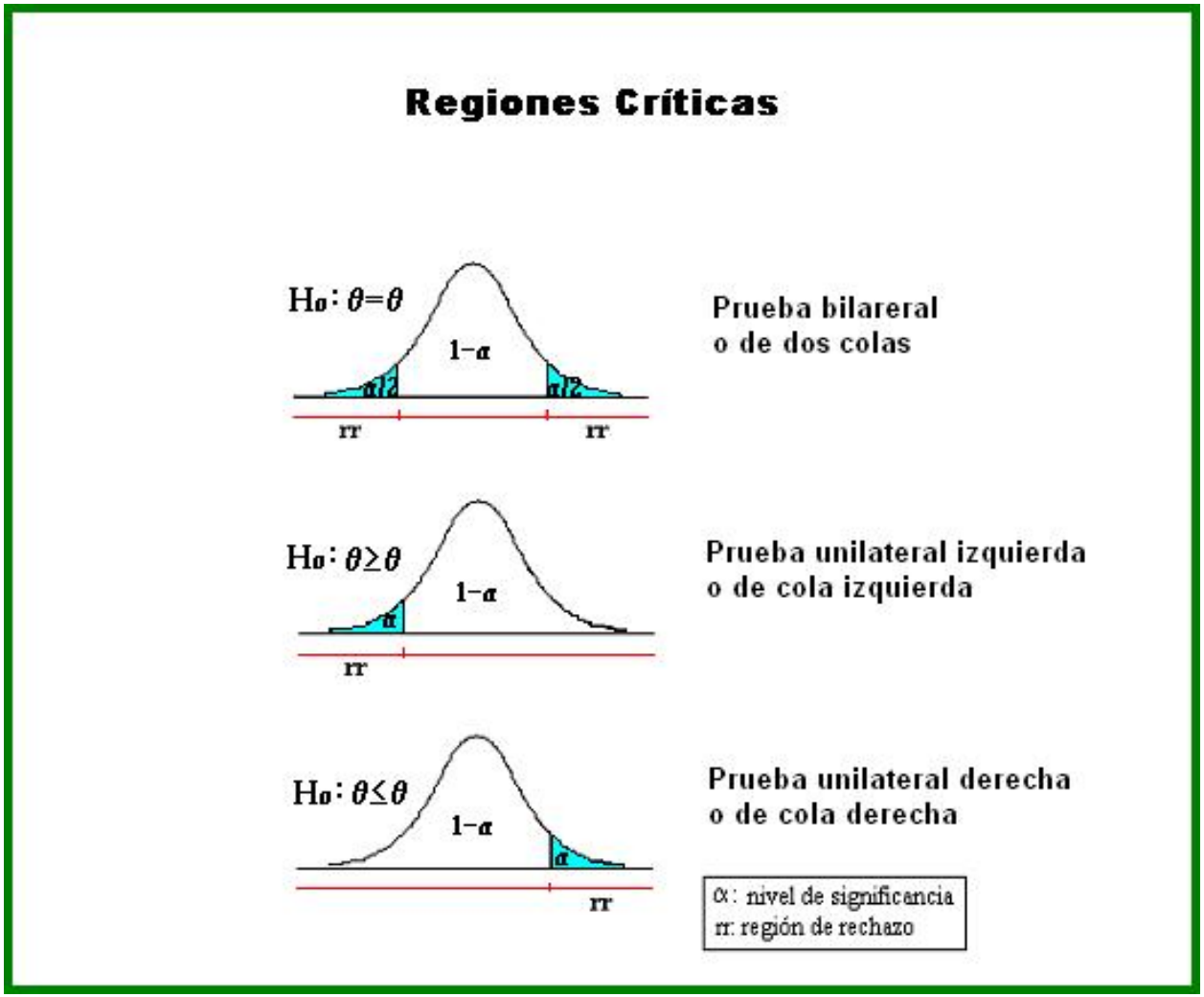
La especificación del contraste sólo afecta a la regla de decisión

Elementos de los contrastes de hipótesis

- Estadístico del contraste: Nos permite decidir si es "probable"que se observen los datos muestrales, suponiendo que la nula sea cierta.
- Regla de decisión, nivel de significación y región crítica.



Elementos de los contrastes de hipótesis



Contrastes de hipótesis sobre la diferencia de medias

Vamos a analizar si existe una brecha salarial entre hombres y mujeres en España. Por lo que queremos responder a la pregunta...
¿Es menor el salario en el caso de las mujeres?

Especificación del contraste

$$H_0 : \mu_M \geq \mu_H \Rightarrow \mu_H - \mu_M \leq 0$$
$$H_a : \mu_M < \mu_H \Rightarrow \mu_H - \mu_M > 0$$

Estadístico de contraste:

$$t = \frac{\bar{x}_H - \bar{x}_M}{\sqrt{\frac{S_H^2}{n_H} + \frac{S_M^2}{n_M}}}$$

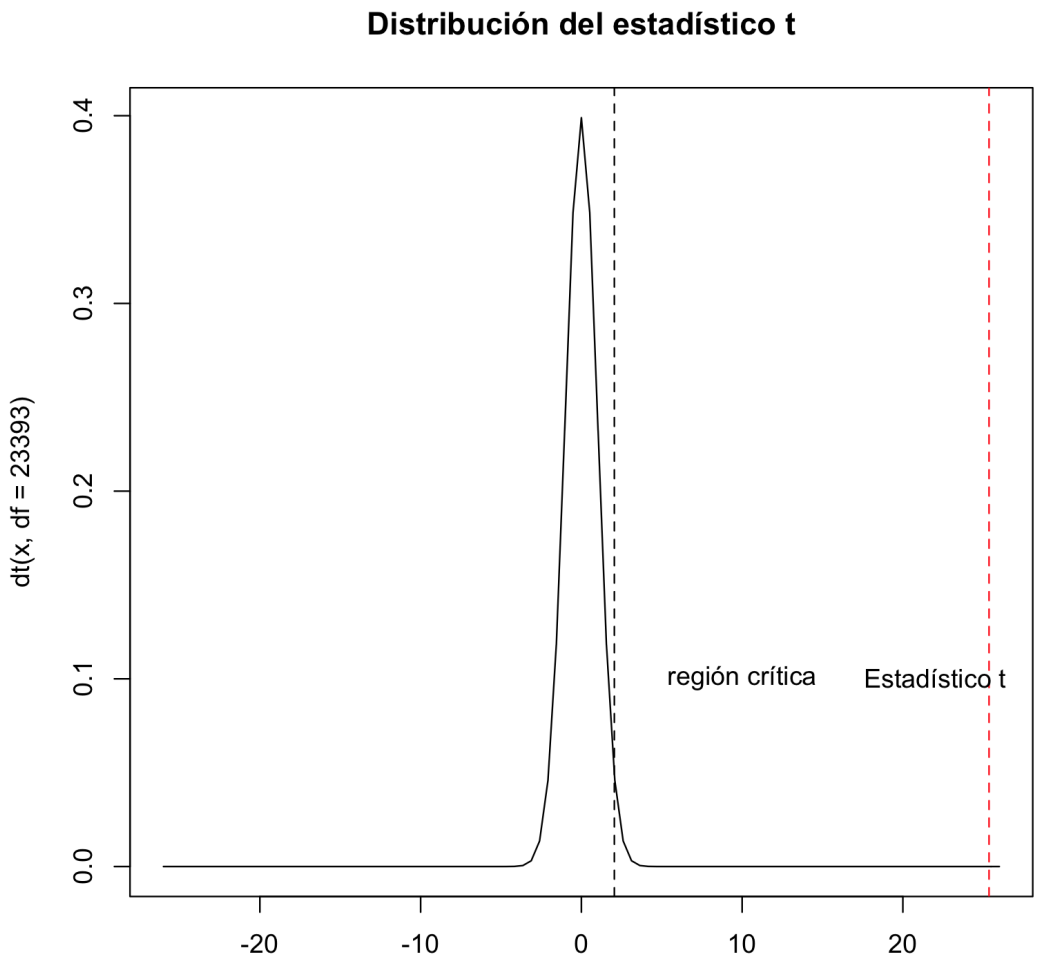
NO CAMBIA!!

Regla de decisión:

$$t > t_{n_H+n_M-2,1-\alpha} \Rightarrow RechazoH0$$

La especificación del contraste sólo afecta a la regla de decisión

Regla de decisión



Contrastes de hipótesis sobre la diferencia de medias

Comprobar antes de realizar el contraste que:

- Independencia de las muestras. Se trata de dos muestras independientes, ya que si eliminamos a uno de los hombres de la muestra no va a afectar en absoluto a la muestra de mujeres. No sería este el caso si la muestra se refiriese a matrimonios, y tuviésemos los hombres por un lado y las mujeres por otro. En ese caso eliminar a uno de los maridos implicaría quitar de la muestra también a su esposa.
- Normalidad. Ambas muestras deben tener un tamaño superior a 30 o en caso de no ser así, las muestras deben distribuirse normalmente.
- Independencia de las observaciones. Las muestras de ambas distribuciones deben ser aleatorias.
- Las varianzas son distintas. El comando que hemos empleado supone por defecto que las varianzas son distintas. Para comprobar dicha hipótesis es recomendable contrastarla, como paso previo a realizar el contraste de igualdad de medias.

Intervalos de confianza para el cociente de varianzas

Sirve para contrastar la hipótesis de igualdad de varianzas.
¿Es igual la dispersión en el salario de las mujeres y de los hombres?

Especificación del contraste

$$H_0 : \sigma_H^2 = \sigma_M^2 \Rightarrow \sigma_H^2/\sigma_M^2 = 1$$
$$H_a : \sigma_H^2 \neq \sigma_M^2 \Rightarrow \sigma_H^2/\sigma_M^2 \neq 1$$

Intervalo de confianza

$$\frac{\sigma_H^2}{\sigma_M^2} \in \left(\frac{S_H^{*2}/S_M^{*2}}{F_{n_H-1,n_M-1;1-\frac{\alpha}{2}}}, \frac{S_H^{*2}/S_M^{*2}}{F_{n_H-1,n_M-1;\frac{\alpha}{2}}} \right)$$

Contrastes de hipótesis sobre la diferencia de medias

El profesor Correy realizó un estudio sobre el efectos del Syntocinon (un medicamento empleado para provocar el parto) en el tiempo transcurrido desde que se rompen aguas hasta el momento del parto. Para ello, durante 12 meses, recogió datos de mujeres que no tomaron el medicamento (grupo de control) y sobre mujeres a las que sí se les administró dentro de un periodo de dos horas desde que rompieron aguas.

- Grupo de control: $n_c = 315; \bar{x}_c = 9{,}43; S_c^2 = 32{,}4616$
- Grupo tratado: $n_t = 301; \bar{x}_t = 9{,}14; S_t^2 = 26{,}2455$

Cuestiones

1. ¿Hay una diferencia significativa entre ambos grupos?
2. ¿Es menor la duración en el caso de las mujeres tratadas?

Contrastes de hipótesis sobre la diferencia de medias

¿Hay una diferencia significativa entre ambos grupos?

Especificación del contraste

$$H_0 : \mu_c = \mu_t$$
$$H_a : \mu_c \neq \mu_t$$

Estadístico de contraste:

$$t = \frac{\bar{x}_c - \bar{x}_t}{\sqrt{\frac{s_c^2}{n_c - 1} + \frac{s_t^2}{n_t - 1}}}$$

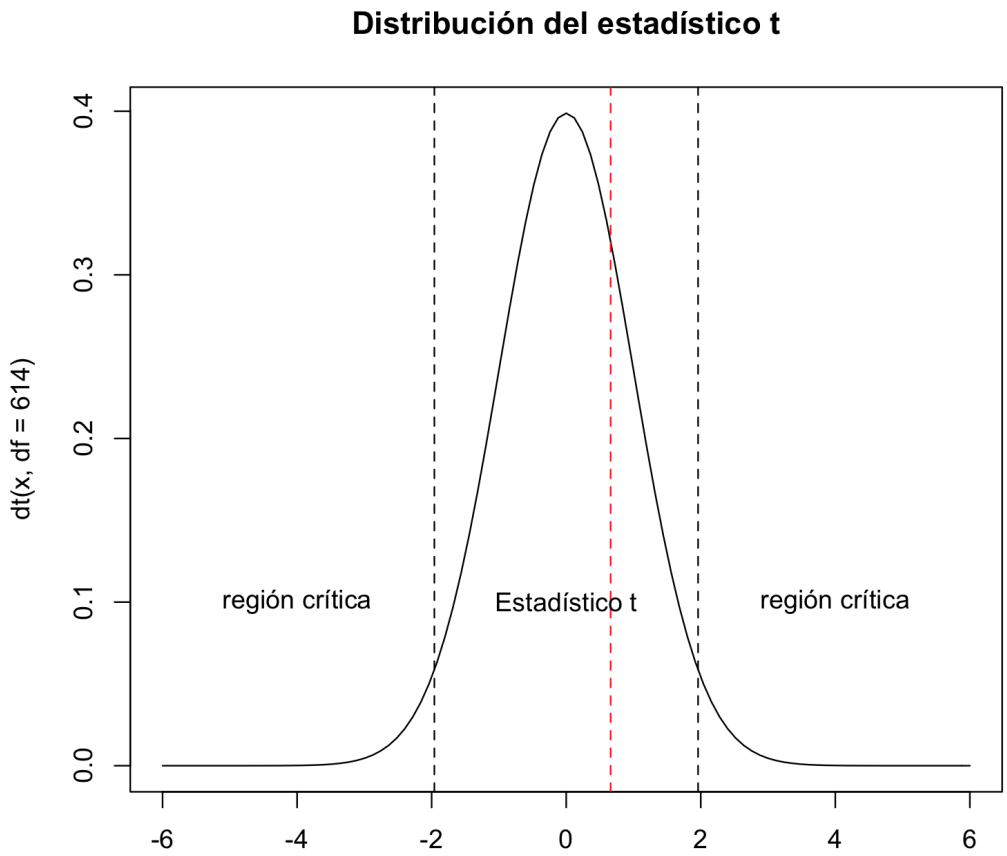
Regla de decisión:

$$|t| > t_{n_c + n_t - 2, 1 - \alpha/2}$$

Este contraste puede realizarse también usando el siguiente intervalo de confianza:

$$\mu_c - \mu_t \in \left(\bar{X}_c - \bar{X}_t \pm t_{n_c + n_t - 2, \alpha/2} \sqrt{\frac{S_c^2}{n_c - 1} + \frac{S_t^2}{n_t - 1}} \right)$$

Regla de decisión



Contrastes de hipótesis sobre la diferencia de medias

¿Es menor la duración en el caso de las mujeres tratadas?

Especificación del contraste

$$H_0 : \mu_c \leq \mu_t$$
$$H_a : \mu_c > \mu_t$$

Estadístico de contraste:

$$t = \frac{\bar{x}_c - \bar{x}_t}{\sqrt{\frac{S_c^2}{n_c - 1} + \frac{S_t^2}{n_t - 1}}}$$

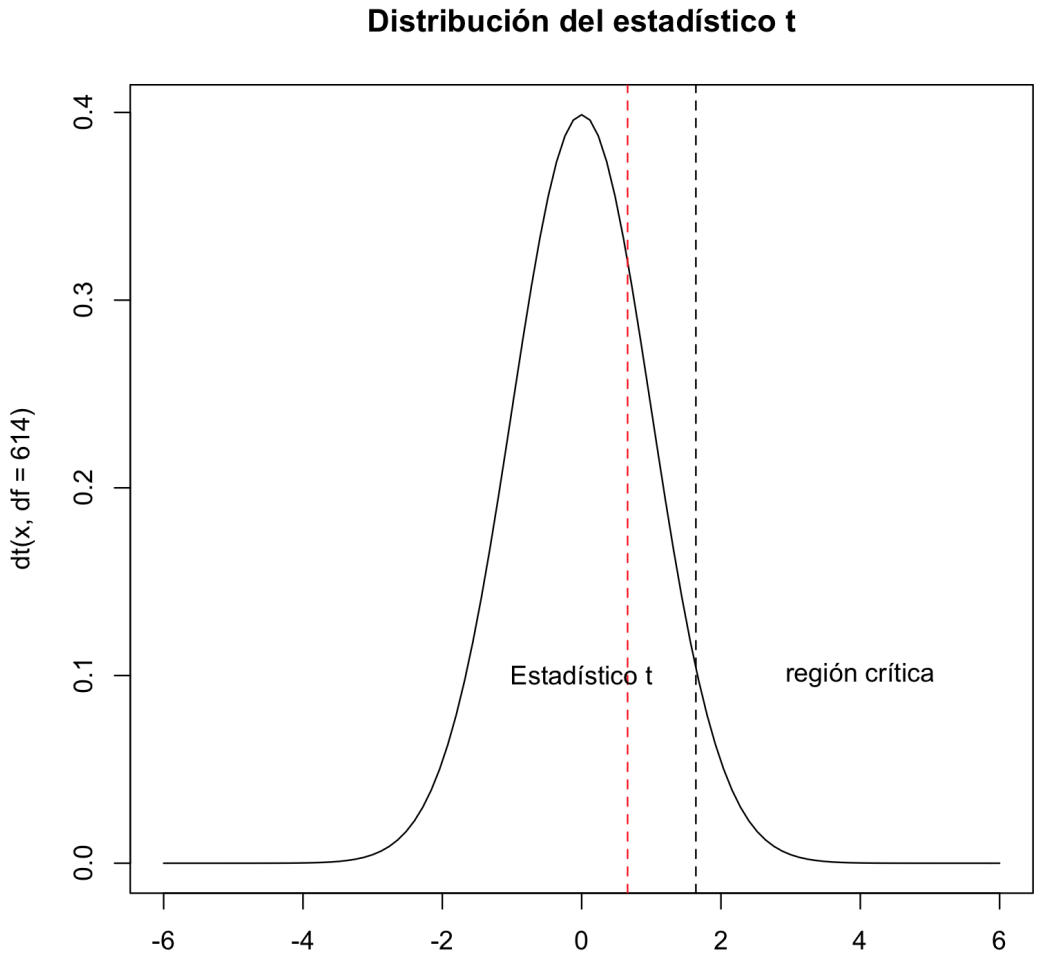
Regla de decisión:

$$t > t_{n_c + n_t - 2, 1-\alpha} \Rightarrow RechazoH0$$

La especificación del contraste sólo afecta a la regla de decisión

Este contraste NO puede realizarse usando intervalos de confianza.

Regla de decisión



Contrastes de hipótesis sobre la diferencia de proporciones

Previo a las elecciones del 20D se realizaron varios sondeos en los que se recogían datos sobre la intención de voto. Las siguientes páginas web presentan los resultados de dos de ellas:

- Encuesta realizada por GAD3 para ABC
- Encuesta presentada por el País

Cuestiones

1. ¿Fueron capaces ambas encuestas de predecir el porcentaje final de votos del PP? (resultado 28,72 %)
2. ¿Asignan ambas encuestas un porcentaje de votos estadísticamente diferente al PP?
3. ¿Es mayor el porcentaje de votos que espera obtener el PP en la encuesta realizada para el ABC?

Contrastes de hipótesis sobre la proporción

¿Fueron capaces ambas encuestas de predecir el porcentaje final de votos del PP?

Especificación del contraste

$$H_0 : pPP_{ABC} = 0,2872$$

$$H_a : pPP_{ABC} \neq 0,2872$$

Estadístico de contraste:

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

donde $p_0 = 0,2872$. Regla de decisión:

$$|z| > z_{1-\alpha/2} \Rightarrow RechazoH0$$

La especificación del contraste sólo afecta a la regla de decisión

Contrastes de hipótesis sobre la diferencia de proporciones

¿Asignan ambas encuestas un porcentaje de votos estadísticamente diferente al PP?

Especificación del contraste

$$H_0 : pPP_{ABC} = pPP_{Pais}$$

$$H_a : pPP_{ABC} \neq pPP_{Pais}$$

Estadístico de contraste (denotando $pPP_{ABC} = p_1, pPP_{Pais} = p_2$):

$$z = \frac{p_1 - p_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p_0(1 - p_0)}}$$

donde $p_0 = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$. Regla de decisión:

$$|z| > z_{c\alpha/2} \Rightarrow RechazoH0$$

La especificación del contraste sólo afecta a la regla de decisión

Contrastes de hipótesis sobre la diferencia de proporciones

- El tamaño de ambas muestras debe ser superior a 30.
- Las muestras han de ser independientes entre sí. Este contraste no sería válido para analizar, por ejemplo, si la diferencia de votos entre PP y PSOE es significativa, ya que las proporciones no son independientes. Una bajada del número de votantes del PP puede tener un efecto positivo sobre los del PSOE, al ser la suma de todos los partidos igual 100 por cien.
- Las muestras deben seleccionarse de forma aleatoria.

Contrastes de hipótesis sobre la diferencia de proporciones

¿Es mayor el porcentaje de votos que espera obtener el PP en la encuesta realizada para el ABC?

Especificación del contraste

$$H_0 : pPP_{ABC} \leq pPP_{Pais}$$
$$H_a : pPP_{ABC} > pPP_{Pais}$$

Estadístico de contraste: $$z = \frac{p_1 - p_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p_0(1-p_0)}}$$ **NO CAMBIA!!**

donde $p_0 = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$. Regla de decisión:

$$z > z_{1-\alpha} \Rightarrow RechazoH0$$

La especificación del contraste sólo afecta a la regla de decisión

Contrastes de hipótesis sobre la diferencia

Anuncio: Contratando los servicios de un determinado centro deportivo es posible perder más 3 kilos en un mes

Peso antes: 85,93,84,87,84,79,85,78,86
Peso después: 78,94,78,87,78,77,87,81,80
¿Se puede considerar publicidad engañosa?

Especificación del contraste

$$H_0 : \bar{D} \leq 3,$$
$$H_a : \bar{D} > 3,$$

donde $\bar{D}$ = peso antes - peso después

Estadístico de contraste:

$$t = \frac{\bar{D} - d}{S_D/\sqrt{n}}$$

Regla de decisión:

$$t > t_{n-1,1-\alpha} \Rightarrow RechazoH0$$

Contrastes de hipótesis sobre la diferencia de proporciones

- La diferencia se distribuye como una distribución normal en muestras menores de 30. Para tamaños de muestra superiores a 30 este supuesto no es necesario.
- Se requiere que la selección de las parejas de la muestra sea aleatoria.

Índice

- Las hipótesis que se contrastan no hacen referencia a los parámetros poblacionales
- Estudiamos tres test basados en el estadístico χ^2
  1. Contraste de bondad de ajuste
  2. Contraste de homogeneidad
  3. Contraste de independencia

Contraste de bondad de ajuste

Empleando ambas encuestas sobre intención de voto en España ¿Podíamos esperar el resultado obtenido el 20 de diciembre de 2015?
PP: 28,72 %; PSOE: 22,02 %; CDS: 13,93 %; Podemos: 20,65 %.

Especificación del contraste

H_0: X sigue la distribución F_0
H_a: X no sigue la distribución F_0

Estadístico de contraste

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \sim \chi^2_{k-1}$$

Regla de decisión:

$$\chi^2 > \chi^2_{k-1,1-\alpha} \Rightarrow RechazoH_0$$

Contraste de homogeneidad de poblaciones

¿Es razonable admitir que las muestras empleadas en las encuestas realizadas por el ABC y el País siguen la misma distribución?

Especificación del contraste

H_0: Las poblaciones son homogéneas
H_a: Las poblaciones no son homogéneas

Estadístico de contraste:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(k-1)(p-1)},$$

donde p es el número de poblaciones.
Regla de decisión:

$$\chi^2 > \chi^2_{(k-1)(p-1),1-\alpha} \Rightarrow RechazoH_0$$

Contraste de independencia

¿Afecta la educación al nivel de felicidad?¿Son variables independientes?

En la Encuesta de Condiciones de Vida, encontramos dos variables que nos permiten medir ambos fenómenos:
Felicidad: Grado de satisfacción con su vida en la actualidad (k= 11 categorías)
Educación: Nivel de los estudios terminados (h= 6 categorías)

Especificación del contraste

H_0: Las características son independientes
H_a: Las características no son independientes

Estadístico de contraste $$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(k-1)(h-1)},$$

donde $e_{ij} = n_{i.} * n_{.j}/n$
Regla de decisión: $\chi^2 > \chi^2_{(k-1)(h-1),1-\alpha} \Rightarrow RechazoH_0$

Tabla de frecuencias

$$\frac{1245 * 326}{25561} = 37?$$

LS\ED	0	1	2	3	4	5	Total
0	37	127	89	41	0	32	326
1	20	59	36	18	1	20	155
2	43	127	96	46	1	40	355
3	62	208	180	90	0	86	629
4	78	373	262	143	3	136	999
5	254	1136	927	545	7	487	3361
6	191	978	915	636	7	645	3378
7	207	1077	1300	1121	13	1378	5103
8	200	1256	1643	1428	15	1945	6495
9	77	453	621	699	4	903	2766
10	76	405	550	425	5	523	1994
Total	1245	6200	6621	5195	60	6200	25561

Regresión lineal simple

¿Cómo afecta la educación a la felicidad?

Cuando el modelo solo incluye una variable dependiente, la técnica que vamos a emplear se denomina modelo de regresión lineal simple, cuya especificación es la siguiente:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

donde y_i denota el nivel de felicidad x_i es el nivel de educación del individuo i, i=1,..., 10. β_0 es el término constante y β_1 es el coeficiente asociado a la variable educación, siendo ambos parámetros a estimar. Por último, ϵ es el término de error.