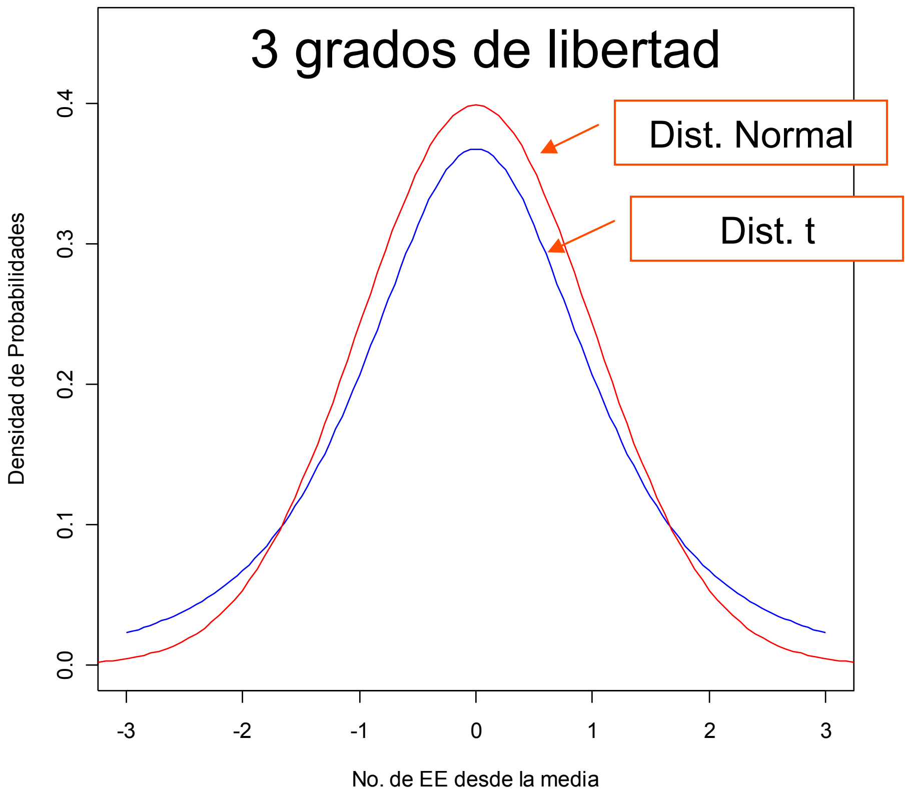
Análisis Estadísticos

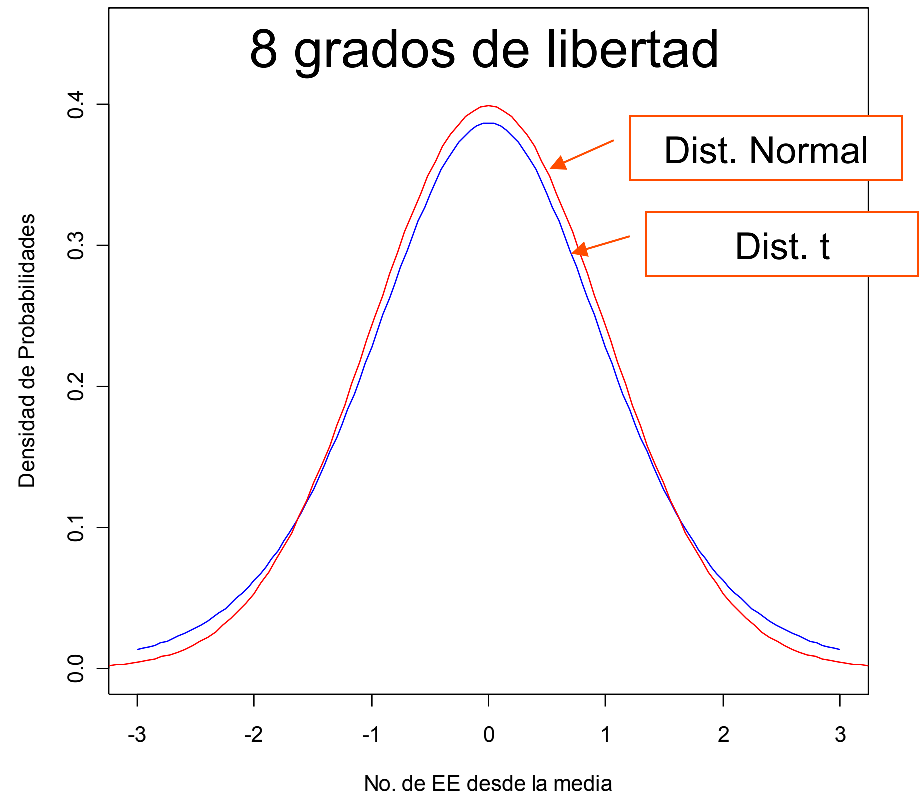
Parte I

Análisis Estadísticos

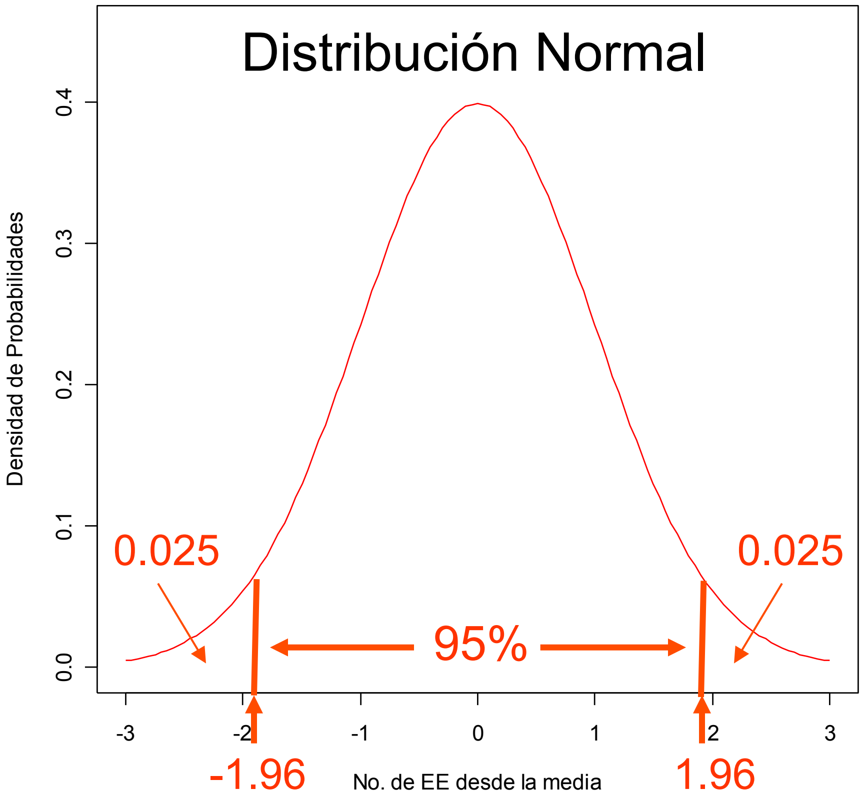
- Análisis para determinar si un parámetro tiene un valor específico.
  - Análisis para determinar si una media tiene un valor específico.
  - Análisis para determinar si una proporción tiene un valor específico
- Análisis para comparar 2 o más estadísticos
  - Comparar 2 medias de muestras independientes
  - Comparar 2 o más medias en un diseño completamente aleatorizado
  - Comparar 2 medias en un diseño pareado
  - Comparar proporciones de muestras independientes

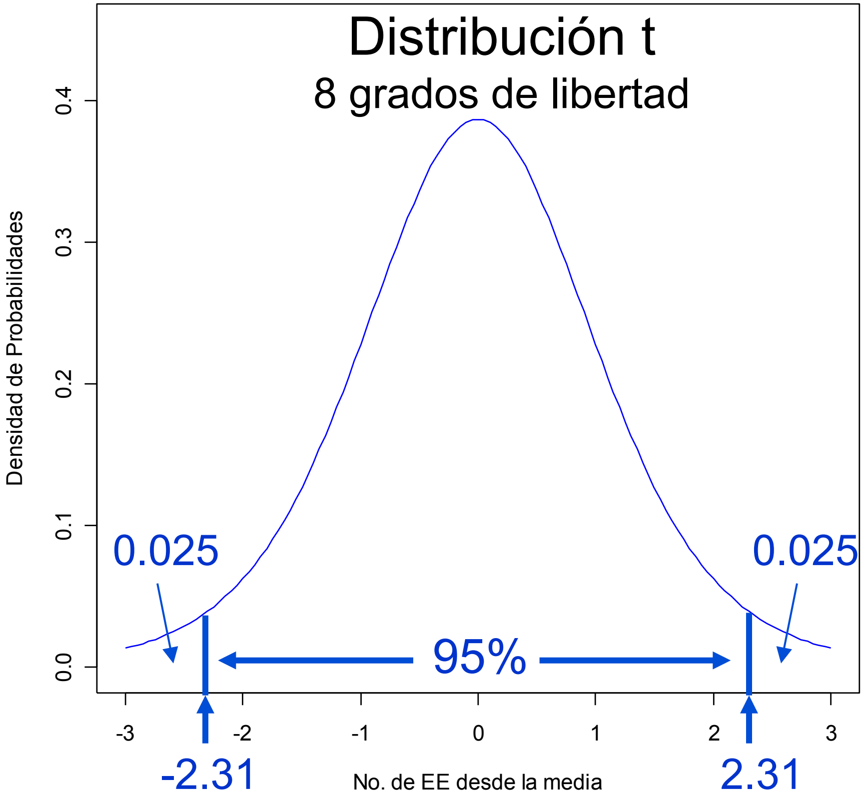
La distribución t





La distribución t

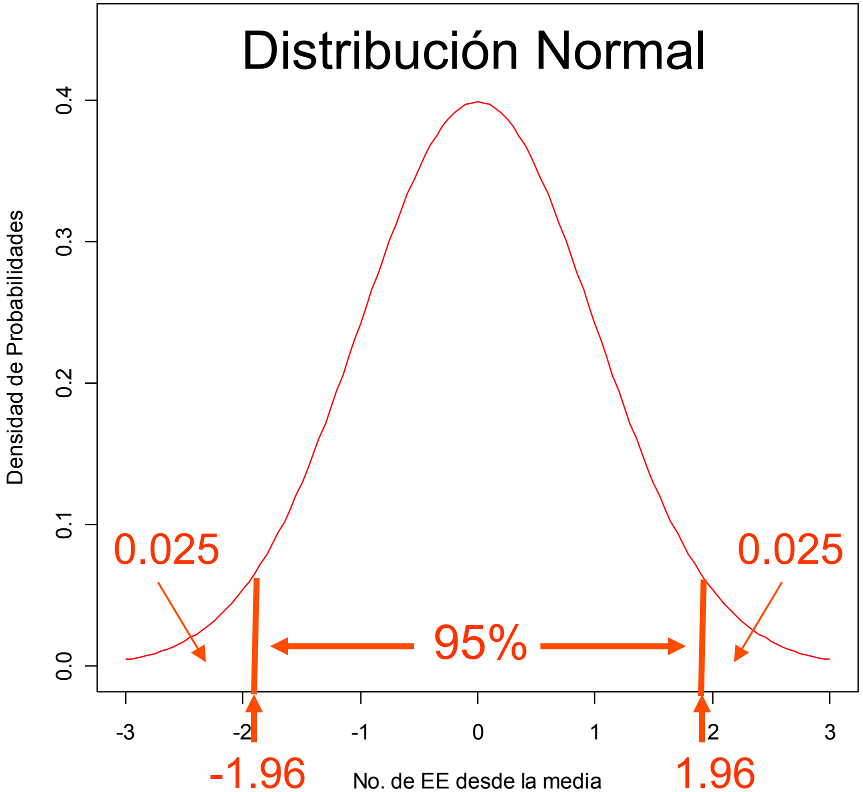


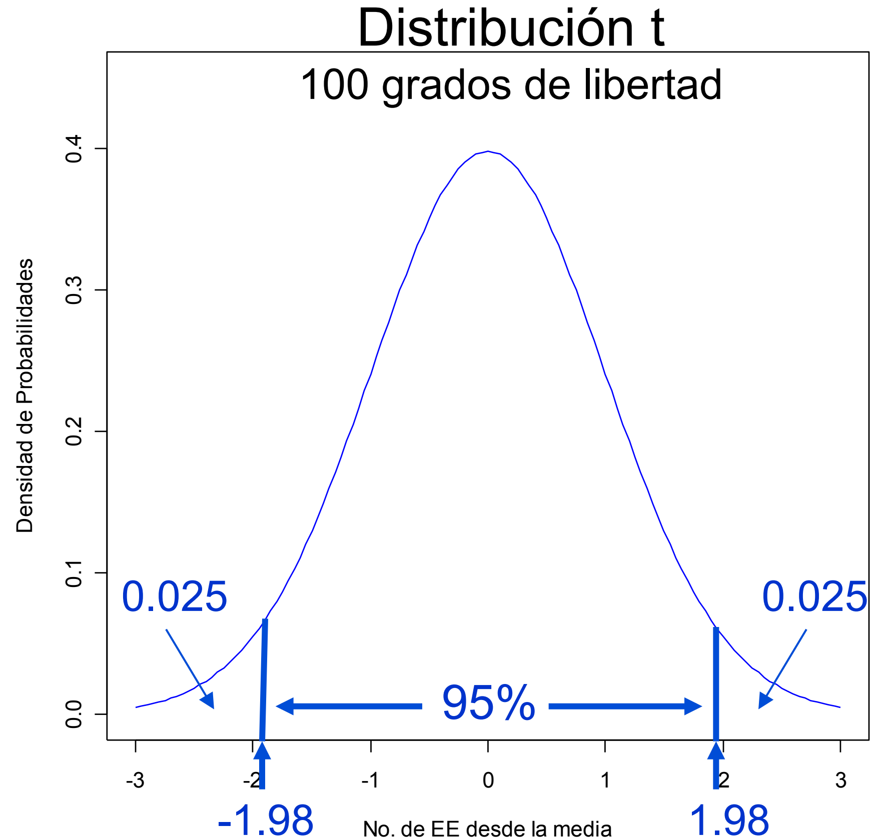


95% de la probabilidad debajo de la curva para una distribución normal se encuentra a 1.96 EE.

95% de la probabilidad debajo de la curva para una distribución de t de 8 grados de libertad se encuentra a 2.31 EE.

La distribución t





95% de la probabilidad debajo de la curva para una distribución normal se encuentra a 1.96 EE.

95% de la probabilidad debajo de la curva para una distribución de t de 100 grados de libertad se encuentra a 1.98 EE.

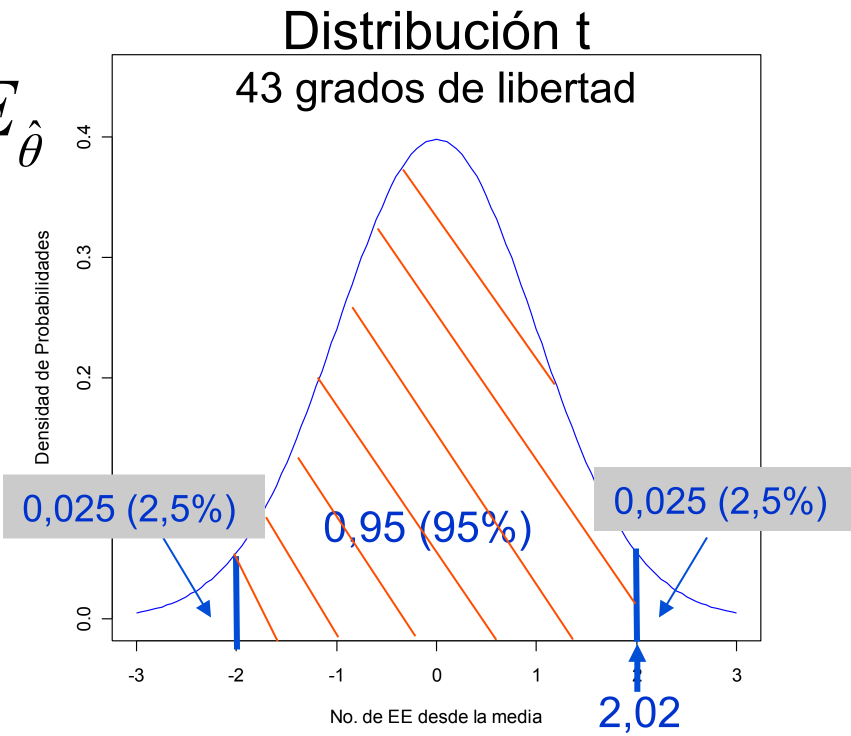
Intervalos de Confianza Normales y la Distribución t

$$\hat{\theta} - t_{(\alpha/2,n-1)} EE_{\hat{\theta}} < \hat{\theta} < \hat{\theta} + t_{(\alpha/2,n-1)} EE_{\hat{\theta}}$$

Para 95%,

alfa (nivel de significancia) = 0,05

alfa/2 = 0,025



Supongamos que tenemos una muestra de 44, por lo tanto, nuestros grados de libertad son 43.

Se debe buscar el valor de t para 43 grados de libertad en una tabla de t (R la tiene) para la probabilidad = 0.025 y se sustituye en la fórmula del intervalo de confianza.

Determinar si una media tiene un valor específico

Ejemplo

- Se quiere determinar si en una población de nutrias el nivel de un contaminante en el torrente sanguíneo es menor a los 300 ppm, nivel que es considerado subletal en nutria.
- Se selecciona al azar una muestra de 25 nutrias y se mide la concentración del contaminante en la sangre.

H0: $\mu = 300$

HA: $\mu < 300$

Prueba de una cola, sólo se pretende si la concentración del contaminante está por debajo de un valor específico.

Se quiere tomar una decisión sobre μ y no sobre $\bar{y}$

Determinar si una media tiene un valor específico

Ejemplo

μ = media poblacional (desconocida) del nivel del contaminante en la sangre de nutrias

$\bar{y}$ = media muestral estimada de las 25 nutrias = 294 ppm

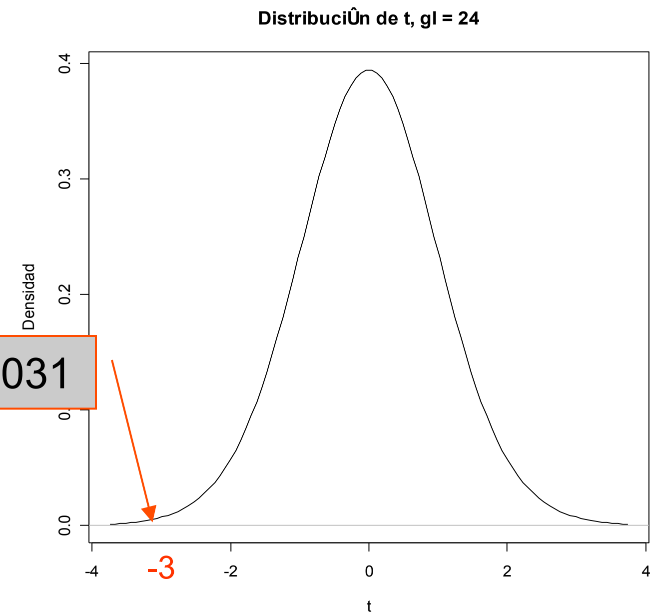
s = desvío estándar muestral = 10 ppm

Aproximación clásica

$$t = \frac{\text{estimado - valor hipotetizado}}{\text{error estándar estimado}}$$

$$t = \frac{294 - 300}{2} = -3$$

P = 0,0031



Se puede concluir que si la hipótesis nula fuera verdadera (si μ = 300) el valor observado de 294 es tan inusual (P = 0,0031) que no se deposita mucha confianza en la hipótesis nula y por lo tanto la rechazamos.

Determinar si una media tiene un valor específico

Ejemplo

μ = media poblacional (desconocida) del nivel del contaminante en la sangre de nutrias

$\bar{y}$ = media muestral estimada de las 25 nutrias = 294 ppm

s = desvío estándar muestral = 10 ppm

Aproximación con Intervalos de Confianza

$$\text{Límite superior} = 298{,}12$$
$$\text{Límite inferior} = 289{,}87$$

El límite superior no supera el valor 300.

Determinar si una proporción tiene un valor específico

Ejemplo

- Se conoce que la tasa de rabia en vampiros es de 20 % bajo condiciones consideradas no epidémicas.
- Se quiere conocer si en una población de vampiros si la tasa de infección es epidémica o no.
- Se muestrean al azar 697 vampiros y 145 se encuentran infectados con el virus de la rabia.

H0: $p_{rabia} = 0{,}2$, $p_{no\ rabia} = 0{,}8$

HA: la tasa de rabia verdadera no es 0,2

Se quiere tomar una decisión sobre p y no sobre $\hat{p}$.

Determinar si una proporción tiene un valor específico

Ejemplo

p = tasa de infección de rabia verdadera (no conocida) en la población de vampiro de interés.

$\hat{p}$ = proporción muestral = 145/697 = 0,208

Aproximación clásica

$$X^2 = \sum \frac{(\text{Conteo observado} - \text{Conteo esperado})^2}{\text{Conteo esperado}}$$

$$X^2 = \sum \frac{(\text{Conteo observado} - np_i)^2}{np_i}$$

$$X^2 = \frac{(145 - 697(0{,}20))^2}{697(0{,}20)} + \frac{(552 - 697(0{,}80))^2}{697(0{,}80)}$$

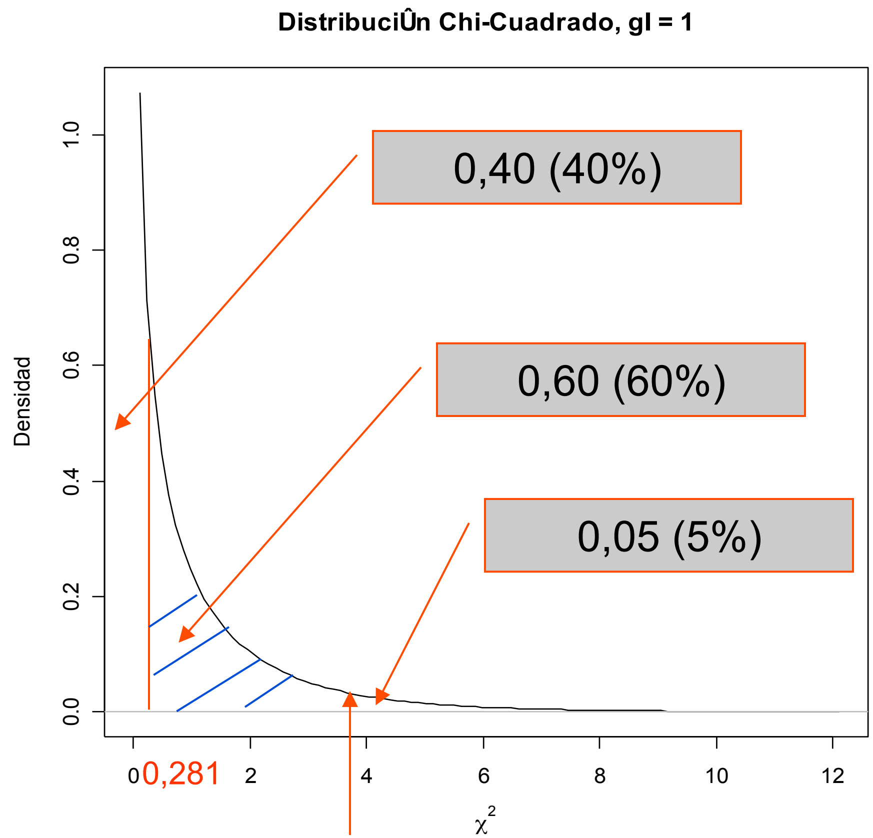
Determinar si una proporción tiene un valor específico

Ejemplo

$$X^2 = \frac{(145-139{,}4)^2}{139{,}4} + \frac{(552-557{,}6)^2}{557{,}6}$$

$$X^2 = 0{,}225 + 0{,}0562 = 0{,}281$$

Valor de p = 0,60



El valor crítico de Chi-cuadrado para 1 grado de libertad y un nivel de significancia alpha de 0,05 es 3,84

Se puede concluir que se falla en rechazar la hipótesis nula y que no hay evidencia de que la tasa de infección de rabia sea epidémica.

Determinar si una proporción tiene un valor específico

Ejemplo

p = tasa de infección de rabia verdadera (no conocida) en la población de vampiro de interés.

$\hat{p}$ = proporción muestral = 145/697 = 0,208

Aproximación con Intervalos de Confianza

$$\text{Error estándar de } \hat{p} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Error estándar de } \hat{p} = \sqrt{\frac{0{,}208(1-0{,}208)}{697}} = 0{,}015$$

Intervalo de confianza

$$\hat{p} \pm 2 \times \text{error estándar de } \hat{p}$$

$$0{,}208 \pm 2 \times 0{,}015 = 0{,}208 \pm 0{,}03 = (0{,}178 \rightarrow 0{,}238) = 17{,}8\% \rightarrow 23{,}8$$

21% (95% IC = 18% - 24%)

Comparar 2 medias de muestras independientes

Comparar 2 medias de un diseño completamente aleatorizado o 2 muestreos independientes

- Este es el caso de la clásica prueba de t de student asumiendo varianzas iguales en ambos grupos.
- No es de interés el valor particular de la media de cada uno de los 2 grupos.
- El interés se focaliza en examinar la diferencia entre las medias de ambos grupos.
- Estamos interesados en diferencias en las 2 direcciones (no sabemos si la media del grupo A es mayor que la del grupo B o viceversa).
- En realidad estamos interesados en conocer si la diferencia es 0 (una diferencia de 0 implica que no hay diferencia entre las medias).

H0: $\mu_A = \mu_B$ o $\mu_A - \mu_B = 0$

HA: $\mu_A \neq \mu_B$ o $\mu_A - \mu_B \neq 0$

Comparar 2 medias de muestras independientes

Comparar 2 medias de un diseño completamente aleatorizado o 2 muestreos independientes

1. Primero se estima la diferencia en medias:

$$\bar{Y}_{\mathrm{A}} - \bar{Y}_{B}$$

2. Luego se calcula el error estándar de la diferencia:

$$s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Para cuando las varianzas no son iguales

$$s_{\bar{y}_1 - \bar{y}_2} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

3. Se calcula un estadístico t para examinar si la diferencia entre medias es 0:

$$t = \frac{\text{Diferencia estimada - Diferencia hipotetizada (0)}}{\text{error estándar estimado de la diferencia}}$$

Se determina la probabilidad del valor de t en una tabla de t.

Comparar 2 medias de muestras independientes

Comparar 2 medias de un diseño completamente aleatorizado o 2 muestreos independientes

- Se asume que ambos grupos tienen varianzas iguales (si no es así entonces se debe usar, por ejemplo, una prueba de Welch).
- No existen observaciones atípicas o inusuales.
- Los datos fueron colectados con un diseño completamente aleatorizado y en muestreos independientes.
- Si la hipótesis nula no es rechazada, entonces se debe considerar el poder de la prueba y el tamaño de la muestra.
- No es necesario tener tamaños de muestras iguales en ambos grupos, pero la potencia de la prueba es maximizada cuando se tienen tamaños de muestras iguales en ambos grupos.
- Un intervalo de confianza es mucho más informativo que una prueba de t interpretada por el valor de P.

Comparar 2 medias de muestras independientes

Comparar 2 medias de un diseño completamente aleatorizado o 2 muestreos independientes

Aproximación alternativa: cálculo del intervalo de confianza

- Se puede examinar la diferencia entre medias mediante el cálculo de un intervalo de confianza.

$$IC = (\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}(s_{\bar{y}_1 - \bar{y}_2})$$

Comparar 2 medias de muestras independientes

Comparar 2 medias de un diseño completamente aleatorizado o 2 muestreos independientes

Ejemplo

- Diferencia entre la densidad media de escarabajos entre islas con y sin aves.

Mediante prueba de t	Mediante un Intervalo de Confianza
t = -1,551, df = 23, P = 0,06	La diferencia entre medias de densidades de escarabajos es de 6.01 individuos por m^2 en favor de las islas con aves pero con un 95% IC = $\underbrace{-2{,}01 \rightarrow 14{,}03}_{\text{Inconcluso}}$ individuos por m^2

Comparar 2 o más medias de un diseño completamente aleatorizado

- Cuando se comparan 2 o más medias se realiza un análisis de varianza (ANDEVA).
- Un ANDEVA para 2 medias es exactamente una prueba de t de student asumiendo varianzas iguales.
- El nombre formal para una prueba de ANDEVA es: Análisis de varianza de un solo factor para un diseño completamente aleatorizado.
- En realidad el ANDEVA es una prueba de la igualdad de las medias a través de la comparación de la variación de los grupos.

Comparar 2 o más medias de un diseño completamente aleatorizado

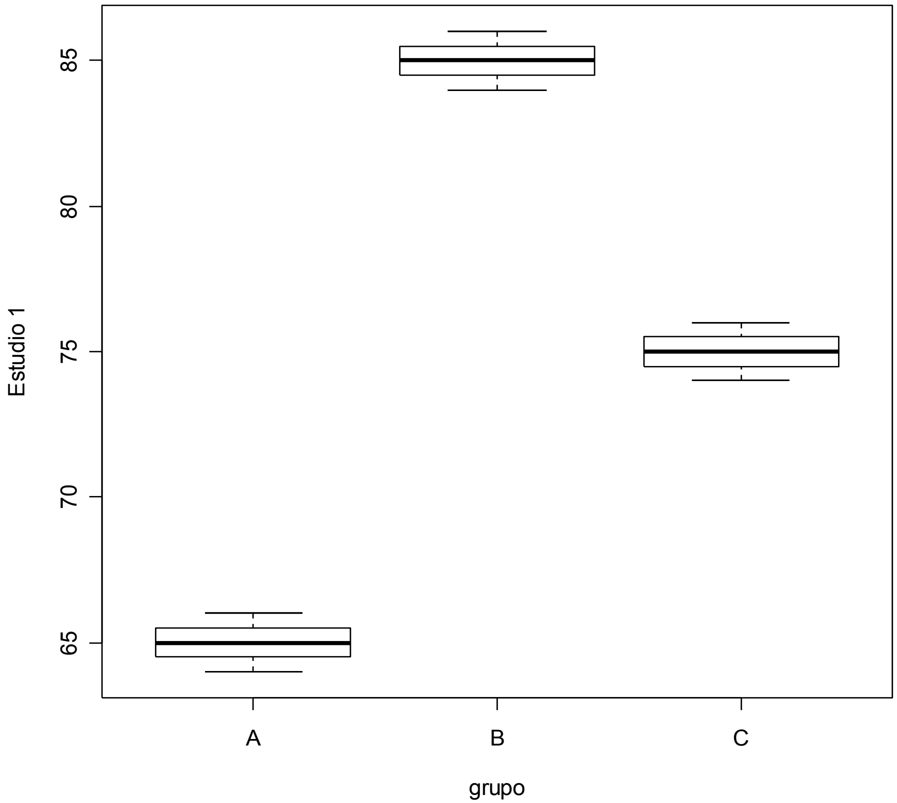
	Estudio 1		
	Grupos		
	A	B	C
	65	84	75
	66	85	76
	64	86	74
Media	65	85	75

Variabilidad

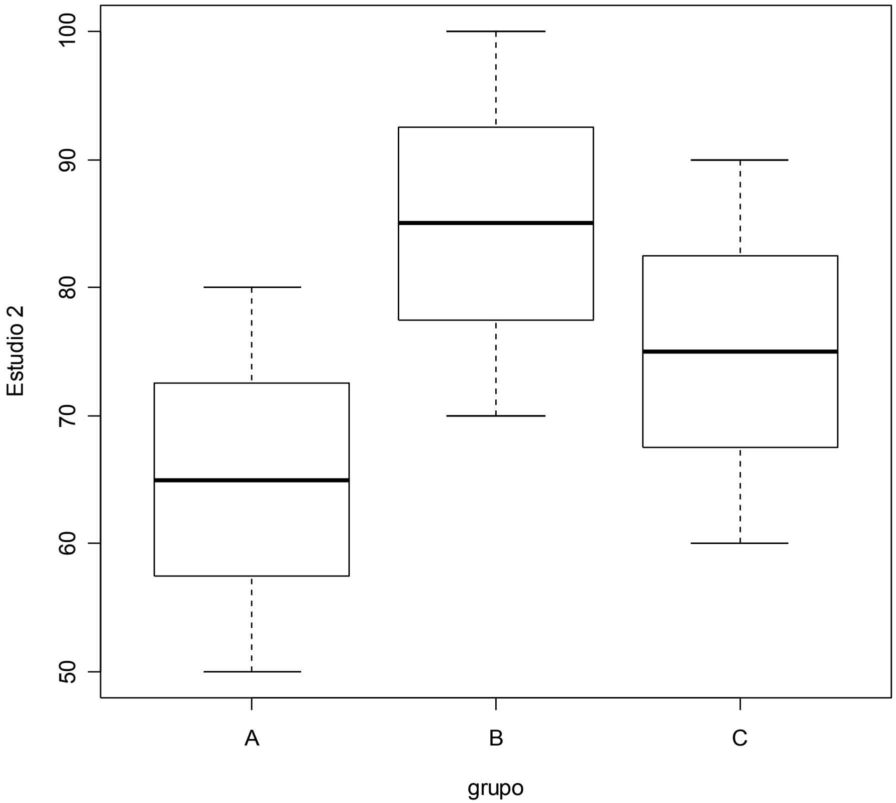
	Estudio 2		
	Grupos		
	A	B	C
	80	100	60
	65	85	75
	50	70	90
Media	65	85	75

Comparar 2 o más medias de un diseño completamente aleatorizado

Estudio 1



Estudio 2



¿Qué estudio posee "mejor" evidencia sobre las diferencias entre las medias de los grupos?

Comparar 2 o más medias de un diseño completamente aleatorizado

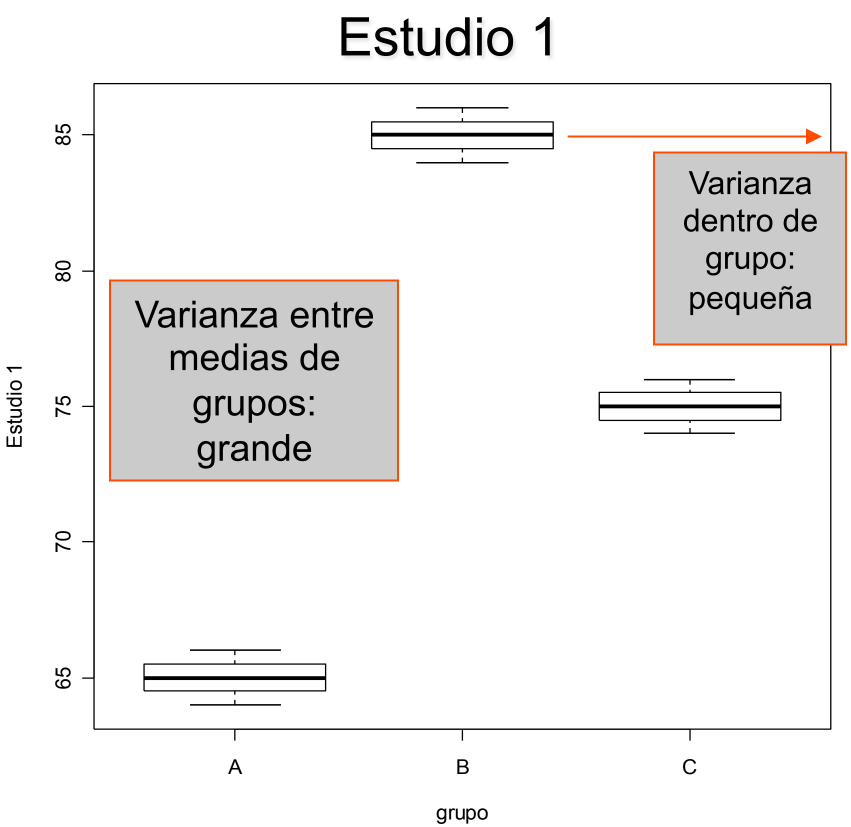
- Aparentemente en el Estudio 1 es más fácil detectar diferencias entre las medias de los grupos. Los resultados son más consistentes.
- El ANDEVA examina las diferencias entre medias a través de la comparación de la variabilidad entre grupos con la variabilidad dentro de cada grupo. Calcula un cociente:

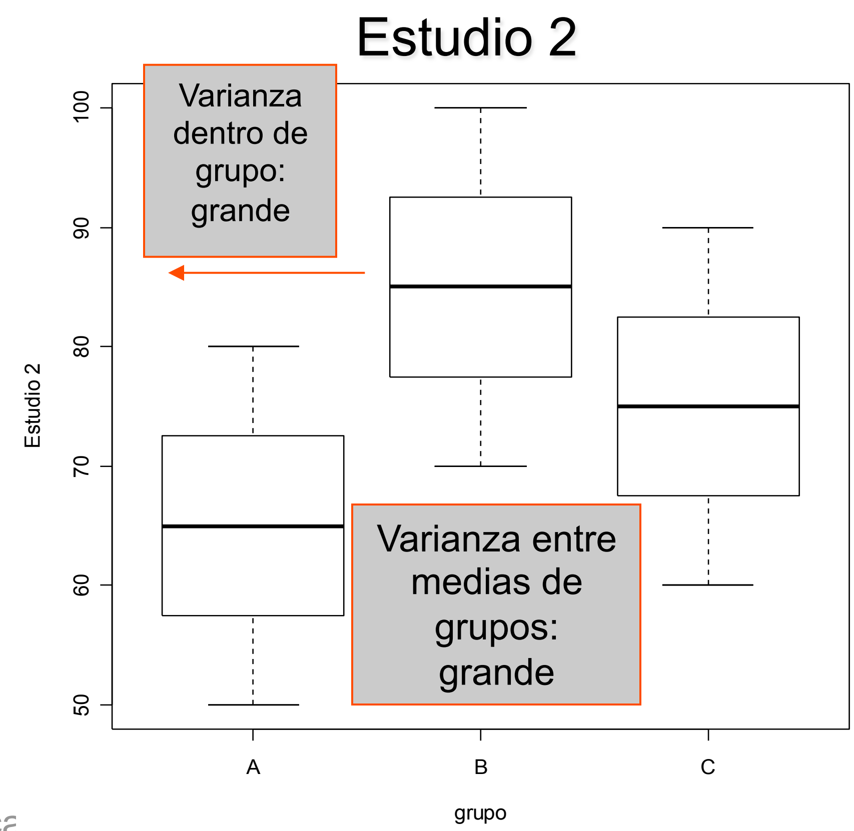
$$\text{Análisis de varianza} = \frac{\text{variación entre grupos}}{\text{variación dentro de grupo}}$$

Si el cociente es grande, existe evidencia para rechazar la hipótesis nula de no diferencia entre las medias

Comparar 2 o más medias de un diseño completamente aleatorizado

- En el Estudio 1, la variabilidad entre la media de los grupos es mucho más grande que la variabilidad de las observaciones dentro de cada grupo.
- En el Estudio 2, la variabilidad entre la media de los grupos no es muy diferente que la variabilidad de las observaciones individuales dentro de cada grupo.





Comparar 2 o más medias de un diseño completamente aleatorizado

$$\text{Análisis de varianza} = \frac{\text{variación entre grupos}}{\text{variación dentro de grupo}}$$

Estudio 1

Varianza entre medias de grupos

Varianza dentro de grupo

Estudio 2

Varianza entre medias de grupos

Varianza dentro de grupo

Comparar 2 o más medias de un diseño completamente aleatorizado

Planteo de hipótesis estadísticas en ANDEVA

H0: $\mu_A = \mu_B = \mu_C$

HA: no todas las medias son iguales o al menos una es diferente del resto

Comparar 2 o más medias de un diseño completamente aleatorizado

Análisis de Varianza – Estudio 1					
Fuente	Grados de libertad	Suma de cuadrados	Cuadrado Medio	Cociente F	P
Grupo (variación entre grupos)	**2**	**600**	**300**	**300**	**<0,001**
Error (variación dentro de grupos)	**6**	**6**	**1**		
Total (variación total)	**8**	**606**	**75,750**		

$$SC_{\text{entre grupos}} = \sum n_{(\text{grupo})} (\text{media grupo} - \text{gran media})^2$$

$$SC_{\text{dentro de grupo}} = \sum \sum (\text{valor} - \text{media grupo})^2$$

Análisis de Varianza – Estudio 2					
Fuente	Grados de libertad	Suma de cuadrados	Cuadrado Medio	Cociente F	P
Grupo (variación entre grupos)	**2**	**600**	**300**	**1,3333**	**0,331**
Error (variación dentro de grupos)	**6**	**1350**	**225**		
Total (variación total)	**8**	**1950**	**243,750**		

Comparar 2 o más medias de un diseño completamente aleatorizado

- La suma de los cuadrados (SC) mide la variación presente en los datos.
- La suma de los cuadrados es dividida en 2 partes, 1) la variación entre medias y 2) la variación dentro de grupos.
- El cuadrado medio es el paso intermedio para encontrar el estadístico de la prueba:

$$CM_{\text{Entre grupos}} = \frac{SC_{\text{Entre grupos}}}{gl_{\text{Entre grupos}}}$$

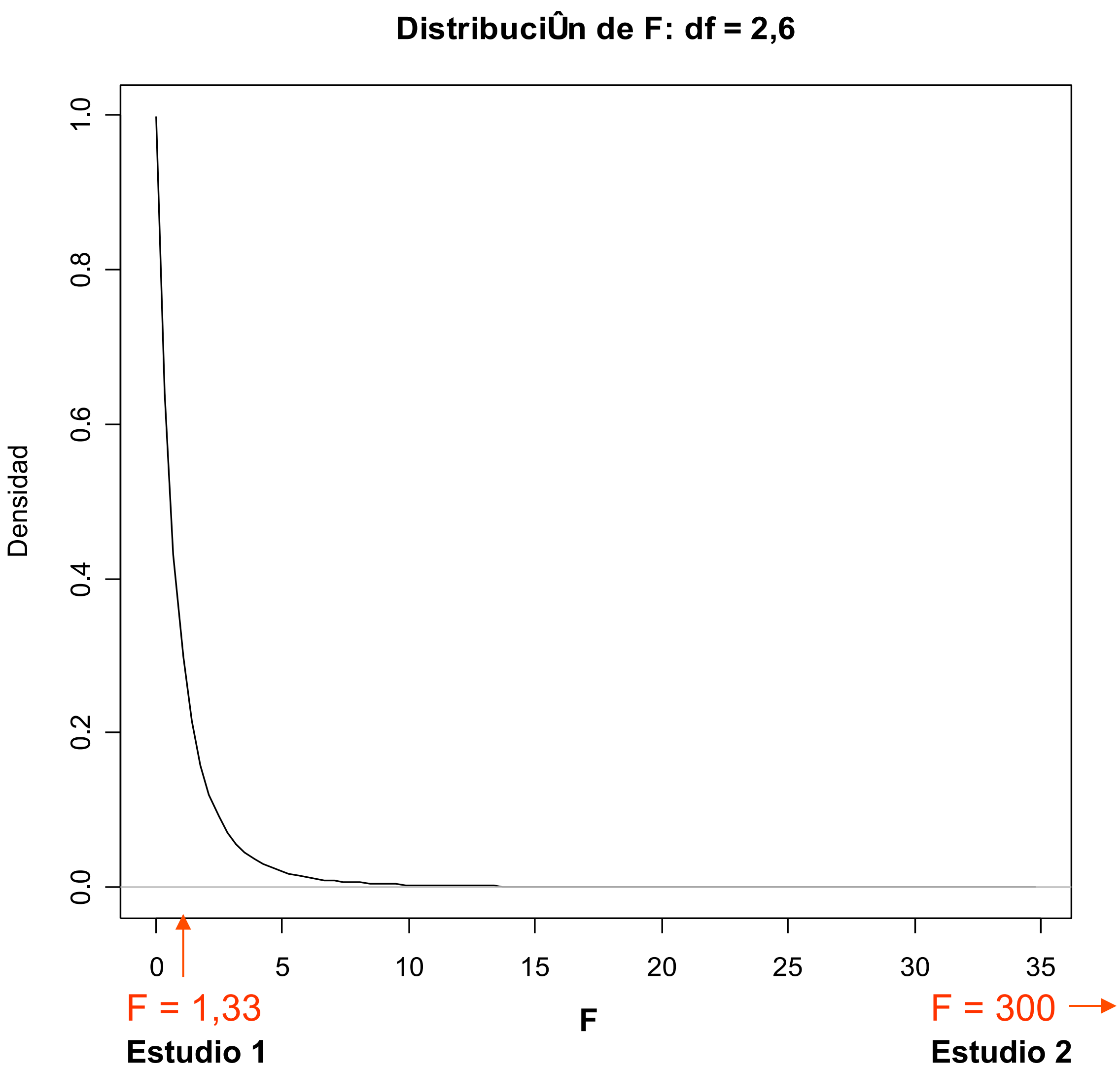
$$CM_{\text{Dentro de grupos}} = \frac{SC_{\text{Dentro de grupos}}}{gl_{\text{Dentro de grupos}}}$$

$$CM_{\text{total}} = \frac{SC_{\text{Total}}}{gl_{\text{Total}}}$$

$$F = \frac{CM_{\text{Entre grupos}}}{CM_{\text{Dentro de grupos}}}$$

$$\text{Análisis de varianza} = \frac{\text{variación entre grupos}}{\text{variación dentro de grupo}}$$

Comparar 2 o más medias de un diseño completamente aleatorizado



Comparar 2 o más medias de un diseño completamente aleatorizado

- El análisis de varianza realizado previamente es únicamente válido si:
  - Los grupos tienen varianzas iguales (similares). Es lo que se conoce como la homogeneidad de varianza.
  - Los grupos y las observaciones son independientes de cada uno.

Homogeneidad de varianza

- Para establecer la homogeneidad de varianza se pueden realizar diferentes pruebas:
  - Prueba de F para 2 varianzas
  - Prueba de Levene
  - Prueba de Bartlett

Comparar 2 o más medias de un diseño completamente aleatorizado

- Cuando se tienen más de 2 medias y se rechaza la hipótesis nula, la pregunta que sigue es, ¿cuáles medias son difieren entre si?
- Para determinar cuales son la medias que difieren se deben usar procedimientos llamados “comparaciones múltiples”.

Comparaciones entre medias (Comparaciones múltiples)

- En los modelos de ANDEVA se hacen comparaciones de las medias de los grupos para determinar el tamaño del efecto entre grupos.
- Sin embargo, como estamos haciendo múltiples comparaciones nos debemos preocupar por dos cosas:
  1. El problema de pruebas múltiples y el incremento del error de Tipo I.
  2. Independencia (ortogonalidad) de los contrastes.

Comparar 2 o más medias de un diseño completamente aleatorizado

Comparaciones múltiples y el incremento del error de Tipo I

- Uno de los problemas más difíciles con las pruebas de hipótesis es la acumulación potencial de errores asociados con la decisión en pruebas múltiples.
- Como el número de pruebas se incrementa, también se incrementa la probabilidad de cometer error de Tipo I entre el conjunto de pruebas.
- La probabilidad de cometer uno o más errores de Tipo I en un conjunto (familia) de pruebas se denomina la tasa familiar de error de Tipo I (Family-wise Type I error rate), también conocida como la tasa experimental de error de Tipo I.
- El incremento de la tasa se da en cualquier situación donde estemos realizando múltiples pruebas de significancia.
- Una familia es definida como una colección de pruebas simultáneas, donde un número de hipótesis son puestas a prueba simultáneamente usando un único conjunto de datos de un experimento o muestreo.

Comparar 2 o más medias de un diseño completamente aleatorizado

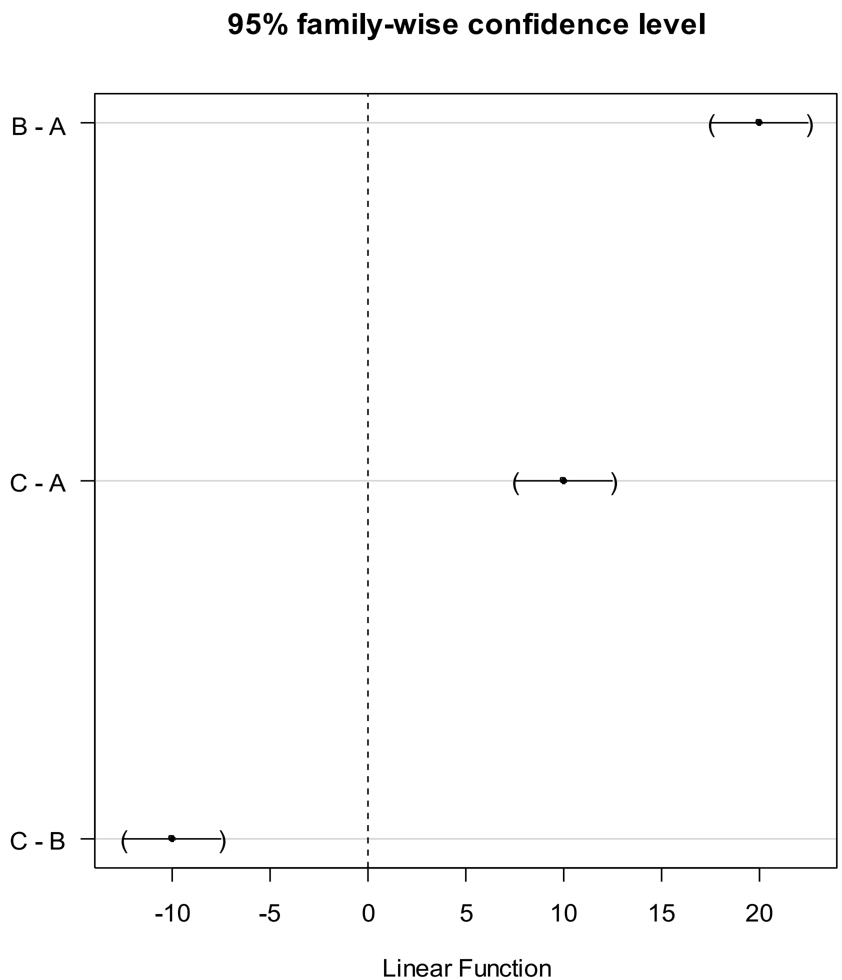
Comparaciones múltiples y el incremento del error de Tipo I

- En algunos casos se comparan todos los posibles pares de medias en una exploración post hoc para encontrar que grupos difieren entre si después de encontrar que el ANDEVA (omnibus test) fue significativo.
- Estas comparaciones no son independientes (mayor número que 1-p).
- El control de la tasa familiar de error Tipo I para un nivel de significancia α se reduce para que cada prueba permanezca en el α escogido.
- Existen 2 aproximaciones para ajustar el nivel de significancia para pruebas múltiples (comparaciones múltiples de medias):
  - Usar pruebas específicas basadas en las distribuciones F o q.
  - Ajustar el valor de P. Es un método más general que puede ser usado para cualquier familia de pruebas.

Comparar 2 o más medias de un diseño completamente aleatorizado

Ejemplo

Prueba HSD de Tukey (Tukey´s HSD test)



Comparar 2 o más medias de un diseño completamente aleatorizado

Ejemplo

Prueba HSD de Tukey (Tukey´s HSD test)

Comparación	Estimado	Límite Inferior	Límite Superior
B - A	20	17,49	22,50
C - A	10	7,49	12,50
C - B	-10	-12,50	-7,49

Comparar 2 medias de un diseño pareado (muestras no independientes)

Comparar 2 medias de un diseño de bloques completamente aleatorizado

- Es conocido como la prueba de t de student pareada.
- Se dice que los datos son pareados porque hay un par de observaciones por cada unidad experimental o de muestreo.
- Se espera que cada par de observaciones estén correlacionadas, porque se midió el mismo individuo o se tomaron del mismo sitio.

La varianza de la diferencia es el promedio de:

$$(y_1 - \mu_1)^2 + (y_2 - \mu_2)^2 - 2\boxed{(y_1 - \mu_1)(y_2 - \mu_2)}$$

Covarianza (correlación)

Cuando la covarianza o correlación es positiva (y grande), reduce la varianza de la diferencia

Comparar 2 medias de un diseño pareado (muestras no independientes)

Comparar 2 medias de un diseño de bloques completamente aleatorizado

- La idea es que la variación entre unidades experimentales o de muestreo puede ser grande comparada a la diferencia en la respuesta, causada por el tratamiento o factor, por lo tanto, se realizan ambos tratamientos o se consideran ambos niveles de un factor en cada unidad experimental o de muestreo.

Comparar 2 medias de un diseño pareado (muestras no independientes)

Comparar 2 medias de un diseño de bloques completamente aleatorizado

Área de acción medidas sobre los mismo individuos en la estación seca y en la estación lluviosa.

Individuo	Área de acción (ha)		Diferencia
	Seca	Lluviosa	
1	31,4	28,1	3,3
2	37,0	37,1	-0,1
3	44,0	40,6	3,4
4	28,8	27,3	1,5
5	59,9	58,4	1,5
6	37,6	38,9	-1,3

Comparar 2 medias de un diseño pareado (muestras no independientes)

Comparar 2 medias de un diseño de bloques completamente aleatorizado

H0: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$ o $\mu_{dif} = 0$

HA: $\mu_1 \neq \mu_2$ o $\mu_1 - \mu_2 \neq 0$ o $\mu_{dif} \neq 0$

$$t = \frac{\text{Diferencia estimada - Diferencia hipotetizada (0)}}{\text{error estándar estimado de la diferencia}}$$

- Los grados de libertad se reducen a la mitad (si se muestrearon 6 individuos, entonces los grados de libertad son 5 (6-1), en vez de 10 (12-2).

Comparar 2 medias de un diseño pareado (muestras no independientes)

Comparar 2 medias de un diseño de bloques completamente aleatorizado

- Se asume que ambos grupos tienen varianzas iguales.
- No existen observaciones atípicas o inusuales.
- Los datos fueron colectados con un diseño pareado.
- Si la hipótesis nula no es rechazada, entonces se debe considerar el poder de la prueba y el tamaño de la muestra.
- El diseño pareado casi siempre es beneficioso. En general, si se tiene información de bloques o con correlación espacial, siempre se debe sacar provecho con un análisis de tipo pareado.
- Un intervalo de confianza es mucho más informativo que una prueba de t interpretada por el valor de P.

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

- En los métodos anteriores se compararon medias.
- En muchos casos la variable “respuesta” puede ser clasificada en clases.
- El interés en este caso es la proporción de la población que “cae” en las diferentes clases.
- Para poner a prueba hipótesis estadísticas relacionadas con proporciones se utilizan pruebas de chi-cuadrado.
- Muchas pruebas estadísticas utilizan la distribución de chi-cuadrado, pero no todas son pruebas de proporciones.
- No todos los análisis de proporciones conducen a una prueba de chi-cuadrado.

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

- Las hipótesis de interés se plantean de 2 maneras diferentes que son equivalentes:

1. Igualdad de proporciones: la hipótesis nula es que la proporción de cada variable "respuesta" es la misma para todos los grupos. La hipótesis alternativa es que las proporciones difieren entre los grupos.
2. Independencia: la hipótesis nula es que la variable respuesta es independiente de los grupos de clasificación. La hipótesis alternativa es que hay una asociación entre la variable respuesta y los grupos de clasificación.

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

- La resumen estadístico básico en el análisis de chi-cuadrado es la tabla de contingencia.
- Las contingencias son todos los eventos que tendrían posibilidad de suceder.
- Una tabla de contingencia muestra los conteos de cuantas veces cada una de las contingencias realmente ocurrió en una muestra determinada.

	Árbol especie A	Árbol especie B
Con hormigas	38	11
Sin hormigas	14	51

- La tabla muestra 4 posibles resultados (contingencias): Árbol especie A con hormigas, Árbol especie B con hormigas, Árbol especie A sin hormigas, Árbol especie B sin hormigas.

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

- El siguiente paso es escoger un modelo que prediga las frecuencias esperadas.
- Hay muchos modelos que se podrían escoger, pero un modelo simple es que la presencia de hormigas es independiente de la especie de árbol (Hipótesis de Independencia).

	Árbol especie A	Árbol especie B	Total de Filas
Con hormigas	38	11	49
Sin hormigas	14	51	65
Total de Columnas	52	62	114

- La probabilidad de que un árbol de la muestra tenga hormigas es 49/114
- La probabilidad de que un árbol de la muestra sea de la especie A es 52/114

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

- Por ejemplo, se quiere conocer la frecuencia esperada de especies de árboles A con hormigas para compararla con la frecuencia observada.
- Como la hipótesis es de independencia nos permite calcular la probabilidad esperada de ser un árbol de la especie A con hormigas como el producto de las 2 probabilidades.

	Árbol especie A	Árbol especie B	Total de Filas
Con hormigas	49/114 x 52/114	49/114 x 62/114	49
Sin hormigas	65/114 x 52/114	65/114 x 62/114	65
Total de Columnas	52	62	114

- Para calcular la frecuencia esperada se multiplica el producto de las 2 probabilidades x la muestra total (114).
- La frecuencia esperada también se calcula como TF x TC/GT (Total de fila x Total de columna dividido por el Gran Total)

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

	Árbol especie A	Árbol especie B	Total de Filas
Con hormigas	38 (22,35)	11 (26,65)	49 (49)
Sin hormigas	14 (29,65)	51 (35,35)	65 (65)
Total de Columnas	52 (52)	62 (62)	114 (114)

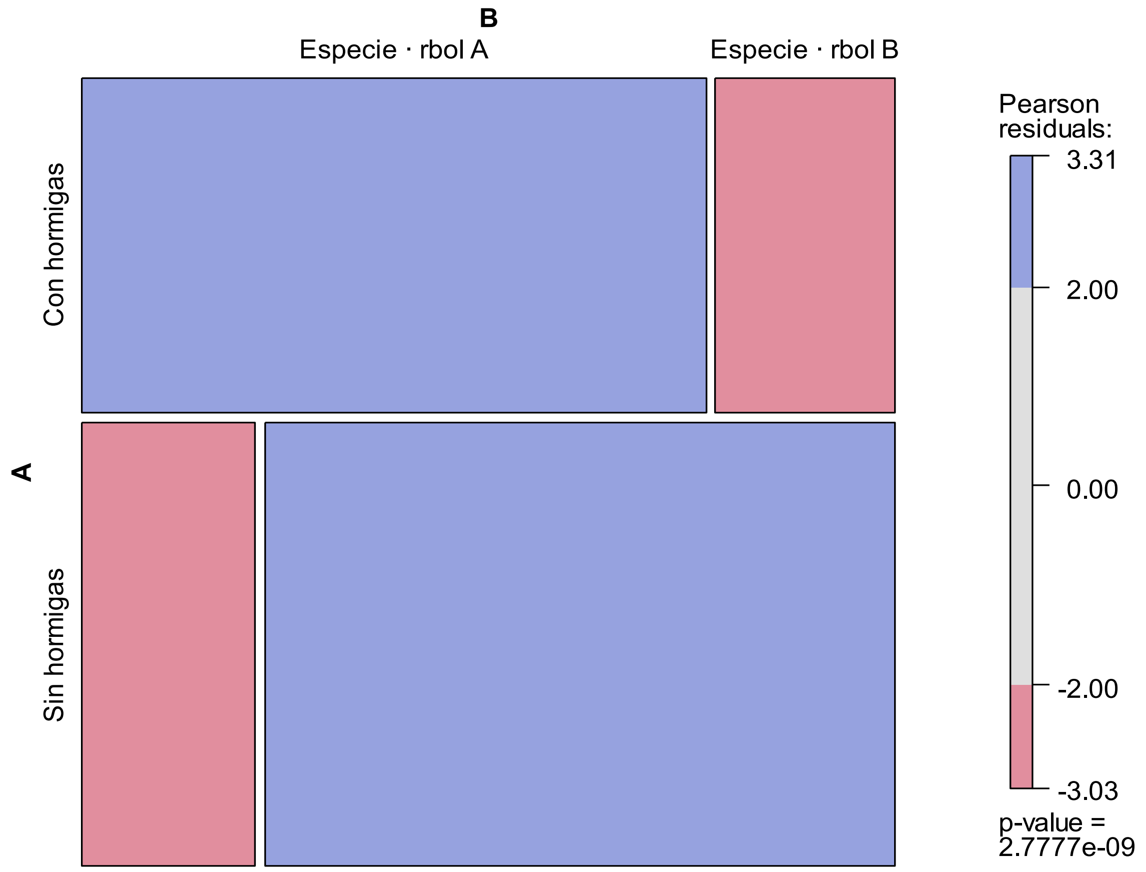
H0 = el resultado (con hormigas o sin hormigas) es independiente de la especie de árbol.

HA = el resultado (con hormigas o sin hormigas) no es independiente de la especie de árbol.

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

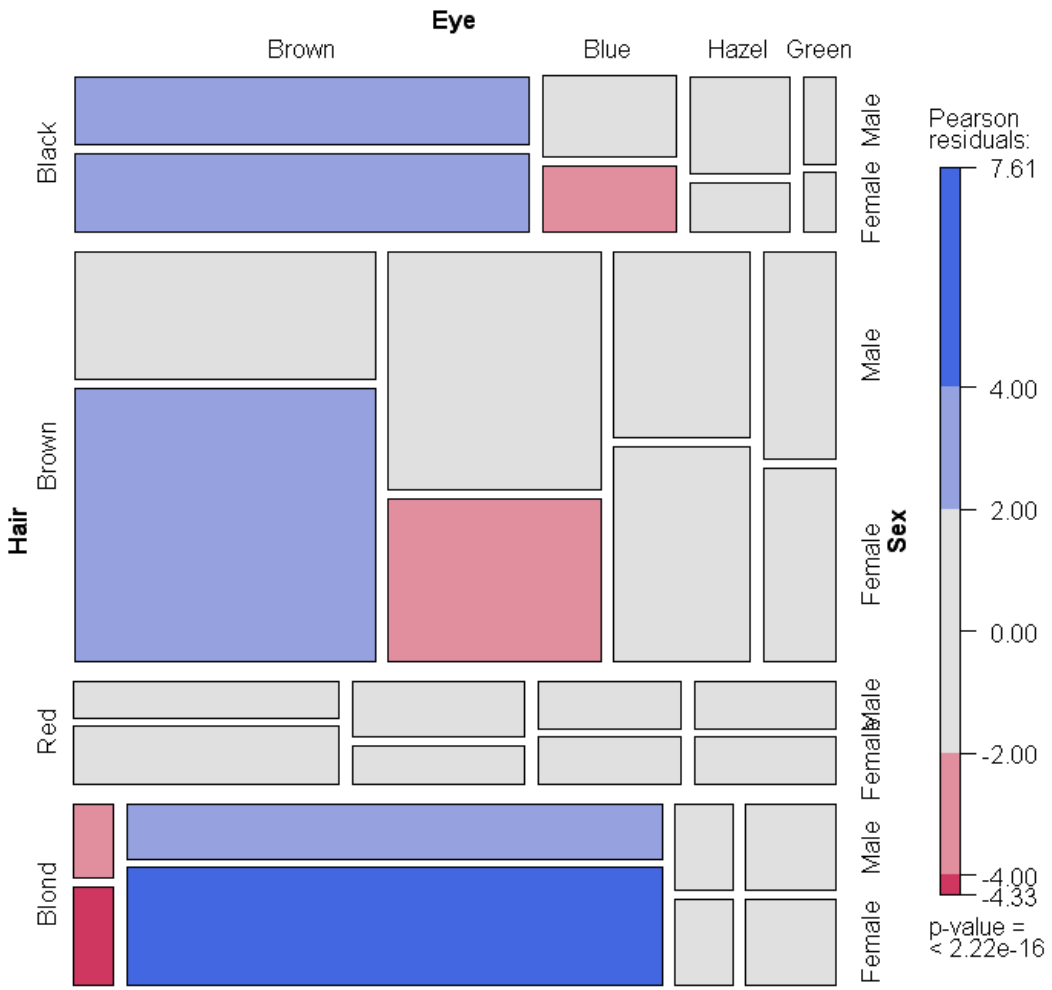
El gráfico de las tablas de contingencia es el gráfico de mosaico.



Comparar proporciones de muestras independientes

Gráfico de mosaico en R

El gráfico de las tablas de contingencia es el gráfico de mosaico.



Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

La estadístico de la prueba se obtiene mediante:

- El chi-cuadrado de Pearson.
- El chi-cuadrado de la razón de Verosimilitud.
- La prueba exacta de Fisher.

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

El chi-cuadrado de Pearson

Frecuencia Observada

Frecuencia Esperada

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia

	O	E	$(O-E)^2$	$(O-E)^2/E$
Especie A Con hormigas	38	22,35	244,92	10,96
Especie B con hormigas	11	26,65	244,92	9,19
Especie A sin hormigas	14	29,65	244,92	8,26
Especie B sin hormigas	51	35,35	244,92	6,93
Total				35,33

Chi-cuadrado = 35,33 para 1 grado de libertad (f-1) x (c-1).

Valor de P < 0,0001

Concluimos que la presencia de las hormigas no es independiente de la especie de árbol (existe una asociación positiva entre las hormigas y la especie de árbol A y una asociación negativa entre las hormigas y la especie de árbol B)

Comparar proporciones de muestras independientes

Prueba de chi-cuadrado de independencia