Tema 6: Inferencia estadística

Contraste de Hipótesis y Estudios Estadísticos

Departamento Matemática Aplicada

Universidad de Málaga

Curso 2020-2021

Inferencia estadística

Cuando hacemos un estudio estadístico de una población de tamaño N, difícilmente podemos estudiar todos sus elementos (resulta costoso, el experimento es destructivo, N puede ser infinito,...) y recurrimos al análisis de una muestra de tamaño $n < N$.

Definición

Llamaremos **muestra** *de tamaño n a un vector* $(X_1, X_2, \ldots, X_n) \in E^n$. *Ese vector tendrá una probabilidad de ser elegido.*

A partir de los resultados obtenidos para esa muestra podemos desear:

1. **Inferencia estadística:** Deducir propiedades de la población completa.
2. **Contraste de hipótesis:** Decidir entre la veracidad o falsedad de una hipótesis.

Ejemplos: Deducir el valor aproximado o un intervalo en el que sea probable que esté la media poblacional, mediana, varianza,..., será un objetivo de la inferencia estadística. Afirmar o rechazar que la media sea mayor que un valor, dos muestras provengan de la misma población, un tratamiento sea mejor que otro, son contrastes de hipótesis.

Técnicas de Muestreo

Definición

Llamamos **Inferencia estadística** *a la parte de la estadística que estudia tanto la forma de selección de los elementos que forman parte de la muestra, como la metodología que lleva a hacer predicciones, referencias, generalizaciones y afirmaciones sobre la población general.*

Definición

Llamaremos **técnica de muestreo** *a la forma de obtención de la muestra a partir de la población general.*

A lo largo del tema siempre usaremos el muestreo aleatorio simple.

Definición

Llamamos **Muestreo Aleatorio Simple** *(m.a.s.), a la técnica de obtención de una muestra en la que todos los elementos de la población* **tienen la misma probabilidad** *de formar parte de la muestra y el muestreo se realiza con reemplazamiento.*

Técnicas de Muestreo

Existen diversas técnicas, pero, para ser eficaces e inferir datos sobre la población, en todas debe ser conocida la probabilidad de que un elemento forma parte de la muestra. Veamos algunos tipos:

- **Muestreo aleatorio simple.**
- **Muestreo estratificado:** Se usa cuando la población no es homogénea, (estudiantes de primaria, secundaria, universidad), entonces el nº de cada clase n_i se extrae en función de su tamaño N_i y variabilidad.
- **Muestreo por conglomerados:** La población está dividida en bloques, la muestra se toma de algunos de los bloques. Por ejemplo, los alumnos de un colegio está dividida en grupos, se eligen algunos grupos y se encuesta a sus alumnos.
- **Muestreo sistemático:** Se obtienen según una regla o fórmula. Por ejemplo: Se ordenan y se toma uno de cada k.

Conceptos y notaciones

- **Tamaño muestral:** Es el número de elementos de la muestra n.
- **Tamaño de la población:** Es el número de elementos de la población N. Puede ser infinito.
- **Estadístico:** Es una función de los valores de la muestra $\xi = f(X_1, X_2, \ldots, X_n)$.Por ejemplo la media muestral.
- **Media muestral:** Es la media de la muestra $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.
- **Media poblacional:** Es la media de la población $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$.
- **Varianza muestral o cuasivarianza de la muestra:** Su valor es:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- **Varianza poblacional:** Su valor es:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Conceptos y notaciones 2

- **Proporción muestral:** Es la proporción de elementos muestrales que verifica una condición A. Se calcula mediante: $\hat{p} = \frac{n_A}{n}$, donde n_A es el número de elementos de la muestra con la característica A.
- **Proporción de la población:** Es la proporción de elementos de la población que verifica una condición A. En el caso de población finita puede ponerse: $p = \frac{N_A}{N}$ donde N_A es el número de elementos de la población con la característica A.

Definición

Llamaremos **estadístico** *a una función que hace corresponder a cada muestra posible un número real.* $f : E^n \to \mathbb{R}$

El valor de un estadístico es a su vez una variable aleatoria. Conocer la distribución que sigue permite obtener más información sobre el significado del mismo para una muestra concreta.
Por ejemplo: Un estadístico es la media muestral $\bar{X} = \frac{\sum_{i=1}^{n} X_n}{n}$ que nos permite inferir el valor de la media poblacional.

Características deseables

De los estimadores $\bar{\xi}$ posibles para un parámetro poblacional ξ, los mejores deben verificar:

Insesgado o centrado: $E(\bar{\xi}) = \xi$

Mínima varianza: De todos los estimadores $\bar{\xi}$, de un parámetro poblacional ξ, será mejor aquel que tenga varianza mínima, pues dará estimaciones más precisas de ξ.

Distribuciones de los estadísticos más usados

Veremos los estimadores centrados y de varianza mínima de los parámetros poblacionales principales para una población $N(\mu, \sigma)$:
Media muestral:

1. La media muestral $\bar{X}$ donde los X_i provienen de una población $N(\mu, \sigma)$, siguen:

$$\bar{X} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

2. La media muestral cuando la varianza poblacional σ es desconocida cumple que:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \rightsquigarrow t_{n-1} \text{ (T-student con n-1 g.d.l.)}$$

Cuando n es grande ($n \geq 30$) usamos la aproximación:

$$\bar{X} \rightsquigarrow N\left(\mu, \frac{S}{\sqrt{n}}\right)$$

Distribuciones estadísticos 2

- **Varianza muestral (cuasivarianza) (S^2):** Su distribución verifica:

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

- **Media muestral para población finita y muestreo sin reposición:**

  Varianza conocida: $\bar{X} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$

  Varianza desconocida y n grande: $\bar{X} \rightsquigarrow N\left(\mu, \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$

Veremos ahora el estimador para la proporción de una población que verifica una condición A:
Proporción muestral: Si n es grande ($n > 30$):

$$\hat{p} \rightsquigarrow N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Estimación

Consideraremos dos tipos de estimación:

1. **Estimación por punto:** A partir de la muestra inferimos el valor más probable del parámetro poblacional deseado.
2. **Estimación por intervalo:** A partir de la muestra inferimos un intervalo I, en el que el parámetro poblacional se encuentra con una probabilidad prefijada.

La estimación por intervalo es más completa que la puntual, pues informa del error que puede estar cometiéndose.

Estimadores puntuales

Los estimadores puntuales centrados y de mínima varianza para los parámetros principales son:

1. **Proporción p:** es la proporción muestral $\hat{p}$.
2. **Media μ:** es la media muestral $\bar{X}$.
3. **Varianza σ^2:** es la varianza muestral S^2.

Ejemplo

Ejemplo

Al realizar un examen, el número de problemas resueltos por un subgrupo de alumnos elegidos al azar, proporciona la siguiente tabla de valores:

Punt.	0	1	2	3	4	5
n_i	10	8	12	9	7	3

Estimar para la población completa formada por 1500 alumnos.

1. *La proporción de alumnos (de los 1500) que contestan al menos dos preguntas.*
2. *La media poblacional μ.*
3. *La varianza poblacional $V = \sigma^2$*

a: El estimador es la proporción muestral $\bar{p} = \frac{31}{49}$

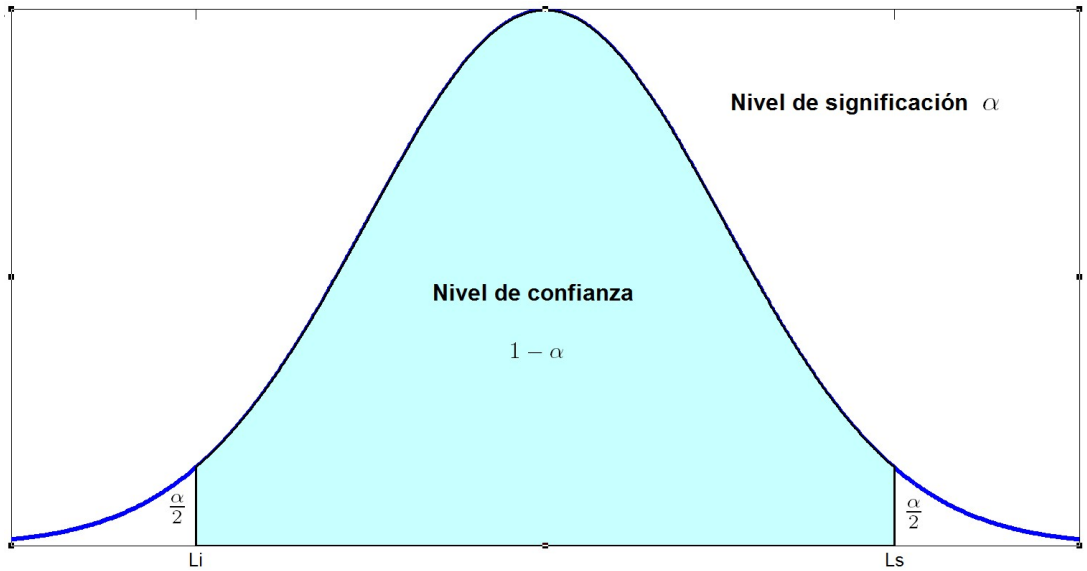
b: El estimador es la media muestral: $\bar{X} = \frac{102}{49}$

c: El estimador es la varianza muestral

$S^2 = \frac{n}{n-1} s_n^2 \approx \frac{49}{48} 2.2791 = 2.3265$

Estimadores por intervalo

Los estimadores por intervalo obtienen un intervalo $I = [L_i, L_s]$ en el que se encuentra el parámetro buscado θ con probabilidad $1 - \alpha$. Es decir, $P(L_i \leq \theta \leq L_s) = 1 - \alpha$ con $0 < \alpha < 1$.
Al valor α se le denomina **nivel de significación** y es la probabilidad de equivocarse al suponer que θ está en el intervalo.
Al valor $1 - \alpha$ se le denomina **nivel de confianza** y es la probabilidad de que θ esté en el intervalo.



Intervalos de confianza

Intervalo de confianza para la media de una normal $\mathrm{N}(\mu, \sigma)$

Varianza Conocida	Varianza desconocida	
	Muestras grandes $n > 30$	Muestras pequeñas $n \leq 30$
$I = \left[\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$	$I = \left[\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$	$I = \left[\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$

Para la varianza σ^2 **de una distribución normal** $\mathrm{N}(\mu, \sigma)$

$$I = \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right]$$

Para el parámetro p **(proporción de una distribución B(n,p))**

$$I = \left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Intervalos de confianza 2

Intervalo de confianza para la diferencia de medias $(\mu_1 - \mu_2)$ de dos distribuciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$

Varianzas	Muestras	Varianzas	Intervalo
Conocidas			$I = \left[(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$
Desconocidas	grandes $n_1 + n_2 > 30$ $n_1 \simeq n_2$		$I = \left[(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$
	Pequeñas $n_1 + n_2 \leq 30$	Iguales	$I = \left[(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$
		Distintas	$I = \left[(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, f}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$

donde: $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ y $f = \dfrac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1+1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2+1}} - 2$ (Welch)

Intervalos de confianza 3

Intervalo de confianza para la razón de varianzas σ_1^2/σ_2^2 de dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$.

$$I = \left[\frac{s_1^2/s_2^2}{F_{\alpha/2;n_1-1,n_2-1}}, \frac{s_1^2/s_2^2}{F_{1-\frac{\alpha}{2};n_1-1,n_2-1}} \right]$$

Intervalo de confianza para la diferencia de parámetros $(p_1 - p_2)$ de dos distribuciones binomiales $B(n_1, p_1)$ y $B(n_2, p_2)$ (muestras grandes $n_1 + n_2 > 30$ y $n_1 \approx n_2$.)

$$I = \left[\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

Ejemplo

Ejemplo

a) Estimar la proporción de fruto que ha sido picado por la mosca de la fruta, si de 150 analizados, se encuentran picados 27.
b) Hallar un intervalo al 95 % de la proporción de picados.

a) $\hat{p} = \frac{n_A}{n} = \frac{27}{150} = 0.18$ (Estimador puntual)
b) Como la proporción dada 95 % es próxima al 100 % nos han dado el nivel de confianza y $\alpha = 0.05$. Sabemos que

$$I = \left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = \left[0.18 \pm 1.96\sqrt{\frac{0.18(0.82)}{150}}\right] = [0.1185, 0.2415].$$

Indicándonos que la proporción se encuentra entre el 11.85 % y el 24.15 % con el 95 % de probabilidad.
El valor $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ puede hallarse mediante las tablas de la distribución normal.

Tipos de contrastes

El otro objetivo principal de la teoría de muestras es confirmar o rechazar, a la vista de los resultados de un experimento, alguna hipótesis sobre la población (generalmente sobre un parámetro θ de la misma). Consideraremos 2 tipos de contrastes:

- **Paramétricos:** Desconocemos el valor del parámetro θ sobre el que se realiza la hipótesis, pero es conocido el tipo de distribución de la población. Por ejemplo sabemos que la población sigue una distribución normal pero no sabemos el valor de su media μ y hacemos la hipotesis de que $\mu = 5$.
- **No paramétricos:** Desconocemos tanto el valor del parámetro objeto de estudio, como la distribución de la población.

Procedimiento

Para realizar un contraste siempre se dan los siguientes pasos:

1. Lo primero es **concretar la hipótesis a contrastar** (una hipótesis cada vez) y **el nivel de significación α del contraste**. Por ejemplo: $\theta = 5$.
2. Esa hipótesis (de igualdad o desigualdad) será la **hipótesis nula** H_0, lo contrario será la **hipótesis alternativa** H_a. (Ej. $H_0 : \theta = 5;\ H_a : \theta \neq 5$).
3. Se define un **estadístico de contraste** $\hat{\theta}$ relacionado con la hipótesis a contrastar y del que debemos conocer su distribución en el muestreo.
4. Suponiendo que H_0 es cierta (hipótesis nula) se calculan dos regiones complementarias:
   - **Región de aceptación:** Región en la que aceptamos H_0 si $\hat{\theta}$ pertenece a la misma.
   - **Región crítica (R):** Región en la que rechazamos la hipótesis nula y aceptamos la alternativa H_a cuando $\hat{\theta}$ pertenece a la misma.
5. A partir de la muestra **se calcula** $\hat{\theta}$ y también el **p-valor** correspondiente y **se toma la decisión final:**
   - Si $\hat{\theta} \in R$ o p-valor $\leq \alpha$ rechazamos H_0 y aceptamos H_a.
   - Si $\hat{\theta} \notin R$ o p-valor $> \alpha$ aceptamos H_0.

Errores al contrastar

Al contrastar podemos cometer dos tipos de error:

- **Error tipo I: Rechazar** H_0 **siendo cierta.** Es lo que se ha denominado nivel de significación α del contraste.

  $$\alpha = \text{P(rechazar } H_0/H_0 \text{ es cierta)} = \text{P(aceptar } H_a/H_0 \text{ es cierta)}$$

  Al nivel de significación más pequeño posible que puede escogerse, para el cual todavía se rechazaría la hipótesis nula con las observaciones actuales de la muestra se le denomina **p-valor**. Cualquier nivel de significación escogido inferior al p-valor comporta aceptar H_0.

- **Error tipo II: Aceptar** H_0 **siendo falsa.** A la probabilidad de cometer ese error se le denomina β. A $1-\beta$ se le denomina **potencia del contraste**.

  $$\beta = \text{P(no rechazar } H_0/H_0 \text{ es falsa)} = \text{P(aceptar } H_0/H_0 \text{ es falsa)}$$

Los errores α y β están relacionados, pues si uno disminuye el otro aumenta. La única forma de disminuir ambos es aumentar el tamaño de la muestra (aumenta el costo del experimento).

Contraste de la media μ de una $N(\mu, \sigma)$.

Regiones críticas:

Varianza	**Muestras**	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$
conocida		$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha$	$\frac{\lvert\bar{x}-\mu_0\rvert}{\sigma/\sqrt{n}} > z_{\alpha/2}$	$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha$
desconocida	grandes $n > 30$	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}} < -z_\alpha$	$\frac{\lvert\bar{x}-\mu_0\rvert}{s/\sqrt{n}} > z_{\alpha/2}$	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}} > z_\alpha$
desconocida	pequeñas $n \leq 30$	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}} < -t_{\alpha,n-1}$	$\frac{\lvert\bar{x}-\mu_0\rvert}{s/\sqrt{n}} > t_{\alpha/2,n-1}$	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}} > t_{\alpha,n-1}$

Contraste de la varianza σ^2 de una $N(\mu, \sigma)$.

Regiones críticas:

$H_0: \sigma^2 \geq \sigma_0^2$ $H_a: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_a: \sigma^2 \neq \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_a: \sigma^2 > \sigma_0^2$
$\frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{1-\alpha,n-1}$	$\frac{(n-1)s^2}{\sigma_0^2} \notin \left[\chi^2_{1-\frac{\alpha}{2},n-1}; \chi^2_{\frac{\alpha}{2},n-1}\right]$	$\frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}$

Contrate del parámetro p (proporción) de una distribución Binomial B(n,p)

Regiones críticas:

$H_0: p \geq p_0$ $H_a: p < p_0$	$H_0: p = p_0$ $H_a: p \neq p_0$	$H_0: p \leq p_0$ $H_a: p > p_0$
$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < -z_\alpha$	$\frac{\lvert\hat{p} - p_0\rvert}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > z_{\alpha/2}$	$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > z_\alpha$

Contraste de hipótesis de la igualdad de medias ($\mu_1 = \mu_2$)

Para dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ de varianzas σ_1^2 y σ_2^2 conocidas.

Regiones críticas:

$H_0: \mu_1 \geq \mu_2$ $H_a: \mu_1 < \mu_2$	$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$	$H_0: \mu_1 \leq \mu_2$ $H_a: \mu_1 > \mu_2$
$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} < -z_\alpha$	$\dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$	$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} > z_\alpha$

Contraste de hipótesis de la igualdad de medias ($\mu_1 = \mu_2$)

Para dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ de varianzas σ_1^2 y σ_2^2 desconocidas para muestras grandes ($n_1 + n_2 > 30$, $n_1 \simeq n_2$)

Regiones críticas:

$H_0 : \mu_1 \geq \mu_2$ $H_a : \mu_1 < \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_a : \mu_1 > \mu_2$
$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} < -z_\alpha$	$\dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} > z_{\alpha/2}$	$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} > z_\alpha$

Contraste de hipótesis de la igualdad de medias ($\mu_1 = \mu_2$)

Para dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ de varianzas σ_1^2 y σ_2^2 desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$) para muestras pequeñas ($n_1 + n_2 \leq 30$, $n_1 \simeq n_2$).

Regiones críticas:

$H_0 : \mu_1 \geq \mu_2$ $H_a : \mu_1 < \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_a : \mu_1 > \mu_2$
$\dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} < -t_{\alpha, n_1+n_2-2}$	$\dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} > t_{\frac{\alpha}{2}, n_1+n_2-2}$	$\dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} > t_{\alpha, n_1+n_2-2}$

donde $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ es la media ponderada de las cuasivarianzas muestrales.

Contraste de hipótesis de la igualdad de medias ($\mu_1 = \mu_2$)

Para dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ de varianzas σ_1^2 y σ_2^2 desconocidas y distintas ($\sigma_1^2 \neq \sigma_2^2$) para muestras pequeñas ($n_1 + n_2 \leq 30$, $n_1 \simeq n_2$)

Regiones críticas:

$H_0 : \mu_1 \geq \mu_2$ $H_a : \mu_1 < \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_a : \mu_1 > \mu_2$
$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} < -t_{\alpha,f}$	$\dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} > t_{\alpha/2,f}$	$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} > t_{\alpha,f}$

donde $f = \dfrac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1+1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2+1}} - 2$ es la aproximación de Welch.

Contraste de hipótesis de la igualdad de varianzas $(\sigma_1^2 = \sigma_2^2)$

Para dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$

Regiones críticas:

$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_a : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_a : \sigma_1^2 \neq \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_a : \sigma_1^2 > \sigma_2^2$
$\frac{s_1^2}{s_2^2} < F_{1-\alpha;n_1-1,n_2-1}$	$\frac{s_1^2}{s_2^2} \notin [F_{1-\frac{\alpha}{2};n_1-1,n_2-1}, F_{\frac{\alpha}{2};n_1-1,n_2-1}]$	$\frac{s_1^2}{s_2^2} > F_{\alpha;n_1-1,n_2-1}]$

Contraste de hipótesis para la igualdad de los parámetros ($p_1 = p_2$)

Para dos distribuciones binomiales $B_1(n_1, p_1)$ y $B_2(n_2, p_2)$ y muestras grandes.

Regiones críticas:

$H_0 : p_1 \geq p_2$ $H_a : p_1 < p_2$	$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$	$H_0 : p_1 \leq p_2$ $H_a : p_1 > p_2$
$E < -z_\alpha$	$\lvert E \rvert > z_\alpha/2$	$E > z_\alpha$

donde $E = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

Contrastes No Paramétricos

En los contrastes parámétricos se presupone que los datos provienen de una distribución conocida, aunque se desconozca algún parámetro de la misma. **En los contrastes no paramétricos no se presupone ningún tipo de distribución subyacente.**
En general, existen contrastes no paramétricos para la toma de decisiones equivalentes a los paramétricos, pero siempre tendrán **menor potencia** que el paramétrico equivalente (menor conocimiento).
Existen contrastes no paramétricos para el valor de la mediana, diferencia de medianas, rachas, signos,...
Nosotros veremos 3 de ellos: Dependencia o independencia, homogeneidad entre muestras y bondad del ajuste.

Ajuste de una distribución a un conjunto de datos

Hemos visto que muchas distribuciones de probabilidad dependen de uno o varios parámetros. Si disponemos de una muestra podemos inferir el valor de los mismos y habremos ajustado una distribución concreta.
Así, dada una muestra de tamaño n para ajustar una distribución:

- **Poisson P(λ):** Estimaremos λ (media) mediante $\bar{x} = \sum_i \frac{x_i}{n}$.
- **Binomial B(N,p) (N conocido):** Bastará con estimar p, mediante la proporción muestral $\hat{p}$.
- **Normal N(μ, σ):** Estimaremos la media μ mediante la media muestral $\bar{x}$ y σ mediante la raíz cuadrada de la varianza muestral s^2.

Para otras distribuciones, usualmente se ajustan sus primeros momentos: media, varianza, . . . , quizás con la única excepción de la uniforme.

Bondad del ajuste

Supongamos que tenemos una población X que puede presentar las modalidades $x_1, x_2, \ldots, x_k$ y suponemos que provienen de una determinada distribución de probabilidad.
Para contrastar si realmente provienen de esa distribución, tomamos una muestra de tamaño n y llamamos:

- o_i a las frecuencias absolutas observadas para cada clase.
- e_i a las frecuencias absolutas esperadas para cada clase si los datos provienen de la distribución supuesta.

Construimos una tabla de contingencia $1 \times k$ mediante:

X	x_1	x_2	$\ldots$	x_i	$\ldots$	x_k
Frec. Observada	o_1	o_2	$\ldots$	o_i	$\ldots$	o_k
Frec. Esperada	e_1	e_2	$\ldots$	e_i	$\ldots$	e_k

Bondad del ajuste 2

Las hipótesis del contraste son:

- H_0 : Los datos provienen de la distribución teórica presupuesta.
- H_a : Los datos no se ajustan a la distribución presupuesta.

Como estadístico de contraste tomamos:

$$\hat{\chi}^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \ldots + \frac{(o_k - e_k)^2}{e_k}$$

que sigue una distribución χ^2 con $k-1$ grados de libertad.
A un nivel de significación α:

- Si $\hat{\chi}^2 < \chi^2_{\alpha;k-1}$ se acepta la hipótesis nula.
- Si $\hat{\chi}^2 \geq \chi^2_{\alpha;k-1}$ se rechaza la hipótesis nula.

Bondad del ajuste 3

Consideraciones:

1. $\hat{\chi}^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$ puede calcularse más fácilmente mediante: $\hat{\chi}^2 = \sum_{i=1}^{k} \frac{o_i^2}{e_i} - n$ donde $\sum_{i=1}^{k} o_i = \sum_{i=1}^{k} e_i = n$.
2. La frecuencia esperada $e_i = np_i$ donde p_i el la probabilidad de la clase x_i.
3. Si hay alguna modalidad con frecuencia esperada $e_i < 5$ debemos agruparla con modalidades contiguas hasta que todas tengan $e_i \geq 5$.
4. Si para obtener el ajuste hemos ajustado m parámetros, entonces los grados de libertad del estadístico son $g = k - m - 1$.
5. Si la distribución a ajustar es continua, determinaremos, si podemos, al menos 5 clases ($k \geq 5$).

Ejemplo bondad del ajuste

Ejemplo

El número de partículas α captadas por un detector Geiger en 1 seg. son:

Número partículas x_i	0	1	2	3	4	5	6 ó más
Número periodos o_i	200	220	150	68	25	10	4

1. *Ajustar una distribución de Poisson.*
2. *Estudiar la bondad del ajuste realizado con α=0.05.*

1: Calculamos la media $\bar{x} = \sum_i \frac{x_i n_i}{n} = \frac{898}{677} \approx 1.3264$, por lo que ajustamos una Poisson de parámetro $\lambda = 1.3264$, es decir: $P(1.3264)$.
2: Calculamos las probabilidades teóricas de cada una de las clases según la Poisson ajustada: $p(\xi = a) = e^{-1.3264} \frac{1.3264^a}{a!}$, obteniendo: $p(\xi = 0) = 0.2654$, $p(\xi = 1) = 0.3521, \ldots, p(\xi = 5) = 0.0091$ y $p(\xi \geq 6) = 1 - \sum_{i=0}^{5} p(\xi = i) = 0.0025$.
Calculamos las frecuencias esperadas: $e_i = np_i = 677p_i$ obteniendo:

x_i	0	1	2	3	4	5	6 ó más
o_i	200	220	150	68	25	10	4
p_i	0.2654	0.3521	0.2335	0.1032	0.0342	0.0091	0.0025
$e_i = np_i$	179.6969	238.3499	158.0737	69.8896	23.1754	6.1480	1.6666

Ejemplo bondad ajuste cont.

Observamos que la frecuencia esperada e_i de la última clase es menor que 5 y debemos agrupar clases, quedando:

x_i	0	1	2	3	4	5 ó más
o_i	200	220	150	68	25	14
p_i	0.2654	0.3521	0.2335	0.1032	0.0342	0.0116
$e_i = np_i$	179.6969	238.3499	158.0737	69.8896	23.1754	7.8146
$\frac{o_i^2}{e_i}$	222.5971	203.0628	142.3387	66.1615	26.9682	25.0814

Obteniendo el estadístico de contraste $\hat{\chi}^2 = \sum_i \frac{o_i^2}{e_i} - 677 = 9.2097$ que se compara con $\chi^2_{0.05;4} = 9.4877$ (g=4, pues hay k=6 clases y se ha estimado la media m=1 (g=k-m-1)).
Como $\hat{\chi}^2 = 9.2097 < 9.4877 = \chi^2_{0.05;4}$, nos quedamos con la hipótesis nula de que el ajuste es bueno.

NOTA: En muchos paquetes estadísticos, al realizar un contraste se devuelve el valor **significación** que corresponde al valor $F(9.2097)$ para la χ^2 con 4 g.d.l. En este caso resulta **sig=0.0561** que se interpreta como que tenemos una probabilidad de equivocarnos al rechazar H_0 del 5.61 %. Al ser mayor que el 5 % nos quedamos con la hipótesis nula. Sin embargo, a un nivel de significación superior a 0.0561 la rechazaríamos.

Contraste de homogeneidad entre muestras

Queremos saber si una muestra es homogénea, es decir, sus elementos provienen de la misma población. Por ejemplo, queremos evitar el efecto de que los encuestadores/observadores usen criterios diferentes, que las mediciones varíen con el tiempo (cansancio,...).

Proceso del contraste:

1. La muestra se descompone en k grupos de tamaños $n_1, n_2, \ldots, n_k$, conteniendo $o_1, o_2, \ldots, o_k$ individuos con la característica A.
2. H_0 : Todas las submuestras provienen de la misma población, nos da una proporción para la característica A de: $p = \frac{o_1+o_2+\ldots+o_k}{n_1+n_2+\ldots+n_k}$
3. Las frecuencias esperadas serán: $e_i = n_i p$ para i=1,2,. . . ,k.
4. Se usa como estadístico de contraste: $\hat{\chi}^2 = \frac{1}{p(1-p)} \sum_{i=1}^{k} \frac{(o_i-e_i)^2}{n_i}$
5. Luego al nivel de significación α:
   - Si $\hat{\chi}^2 < \chi^2_{\alpha;k-1}$ aceptamos la hipótesis nula.
   - Si $\hat{\chi}^2 \geq \chi^2_{\alpha;k-1}$ rechazamos la hipótesis nula.

Ejemplo homogeneidad entre muestras

Ejemplo

Un estudio desea determinar la uniformidad del conocimiento de inglés según el centro de estudio. Para ello, se les realiza un test, obteniéndose los siguientes resultados:

Centro	*A*	*B*	*C*	*D*
Aprobados	30	40	40	35
Presentados	120	200	160	120

Contrastar al 2 % si existen diferencias de conocimiento entre los centros educativos.

Planteamos el contraste:

- H_0 : Los centros son homogéneos respecto al nivel de inglés.
- H_a : Los centros no son homogéneos respecto al nivel de inglés.

Proporción de aprobados: $p = 145/600 = 0.2417$.

Ejemplo homogeneidad entre muestras cont.

Centro	*A*	*B*	*C*	*D*
o_i	30	40	40	35
n_i	120	200	160	120
$e_i = n_i \cdot p$	29	48.3333	38.6666	29
$(o_i - e_i)^2/n_i$	0.0083	0.3472	0.01111	0.3

Estadístico de contraste: $\hat{\chi}^2 = 3.6379$ que se compara con $\chi^2_{0.02,3} = 9.837$.

Como $\hat{\chi}^2 < \chi^2_{0.02,3}$ aceptamos la H_0 de que los centros son uniformes respecto al conocimiento de inglés.

Tablas de contingencia. Contraste de dependencia o independencia.

Queremos contrastar si dos caracteres son dependientes o independientes. Para ello se toma una muestra y se crea una tabla bidimensional (tabla de contingencia) de frecuencias absolutas observadas:

$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_m	Marg. X
x_1	o_{11}	$\dots$	o_{1j}	$\dots$	o_{1m}	$o_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	o_{i1}	$\dots$	o_{ij}	$\dots$	o_{im}	$o_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	o_{k1}	$\dots$	o_{kj}	$\dots$	o_{km}	$o_{k.}$
Marg. Y	$o_{.1}$	$\dots$	$o_{.j}$	$\dots$	$o_{.m}$	n

Consideramos como hipótesis nula (H_0) la independencia de caracteres. En ese supuesto: $p_{ij} = p_{i.}p_{.j} = \frac{o_{i.}}{n}\frac{o_{.j}}{n}$ y $e_{ij} = np_{ij} = \frac{o_{i.}o_{.j}}{n}$ (debe verificarse $e_{ij} > 5$)

El estadístico de contraste será: $\hat{\chi}^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{o_{ij}^2}{e_{ij}} - n$ que sigue una $\chi^2_{(k-1)(m-1)}$.

Así, cuando $\hat{\chi}^2 < \chi^2_{\alpha;(k-1)(m-1)}$ se acepta H_0 y cuando $\hat{\chi}^2 \geq \chi^2_{\alpha;(k-1)(m-1)}$ se rechaza.

Ejemplo independencia entre variables

Ejemplo

Se intenta determinar si existe interacción entre la roya y el moteado del peral (dos enfermedades por hongos). Para ello, se seleccionan 500 parcelas y se confecciona la siguiente tabla:

Roya\Moteado	*Nada*	*Indicios*	*Atacado*
Nada	198	28	62
Indicios	39	6	12
Atacado	105	15	35

Estudiar al 5 % de significación si puede admitirse que el ataque por una u otra enfermedad es independiente entre sí, o por el contrario, el ataque por alguna de ellas se ve favorecido, o repelido, por la presencia de la otra.

Ejemplo independencia entre variables cont.

Planteamos el siguiente contraste:

- H_0 : Las enfermedades son independientes.
- H_a : Las enfermedades son dependientes.

Se construye la tabla de frecuencias esperadas e_{ij} en el caso de que H_0 sea cierta, es decir las enfermedades sean independientes ($e_{ij} = o_{i.}o_{.j}/n$).

e_{ij}	*Nada*	*Indicios*	*Atacado*
Nada	196.992	28.224	62.784
Indicios	38.988	5.586	12.426
Atacado	106.02	15.19	33.79

El estadístico de contraste es $\chi^2 = 0.1175$ que se compara con $\hat{\chi^2}_{0.05,4} = 9.488$. Por tanto, se acepta la hipótesis H_0 de que no existe interacción entre las dos enfermedades.

Estudios estadísticos

- En un estudio estadístico buscamos encontrar relaciones entre diversas variables estadísticas.
- Por ejemplo:
  - El peso y la altura de un individuo.
  - La situación socioeconómica de un alumno y sus resultados en el sistemas escolar.
- Estas relaciones son relaciones probabilísticas, no deterministas.
- La herramienta básica en estos estudios es el Contraste de Hipótesis.

Estudios exploratorios y confirmatorios

- A priori, por razones externas podemos presuponer ciertas hipótesis sobre dichas relaciones y nuestro objetivo es confirmarlas. En dicho caso estamos ante un estudio confirmatorio.
- Pero otras veces, no tenemos claro dichas relaciones y nuestro objetivo es encontrarlas. En dicho caso debemos realizar un estudio exploratorio.
- **Estudio exploratorio:** Estudio en el que buscamos encontrar posibles hipótesis que posteriormente deberán ser comprobadas.
- **Estudio confirmatorio:** Estudio en el que buscamos verificar una hipótesis previamente formulada.

Preguntas de investigación

En los estudios estadísticos confirmatorios buscamos responder **preguntas de investigación**, que relacionan una variable con otra:

- ¿El tabaco produce cáncer?
- ¿El cinturón de seguridad disminuye el número de muertes en accidentes automovilísticos?
- ¿El cambio que hemos realizado en el sistema de producción ha mejorado la rentabilidad?

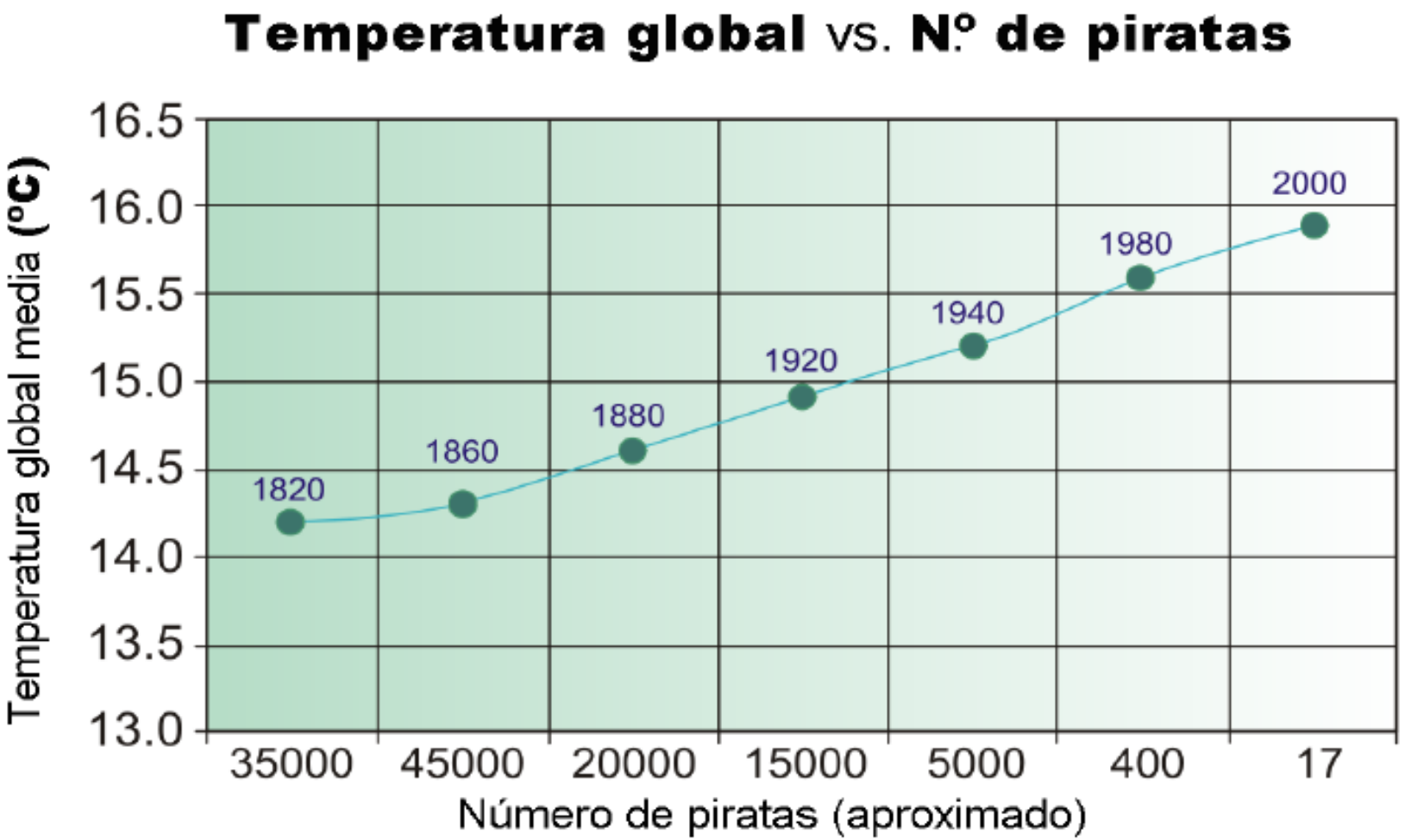
Falacias frecuentes en la realización de estudios estadísticos

- Tanto en los estudios exploratorios como en los estudios confirmatorios la herramienta básica es el contraste de hipótesis.
- Pero, a menudo, se comenten errores en la interpretación de los datos.
- Aquí analizamos dos de las principales falacias lógicas que se comenten con mucha frecuencia:
  - La falacia «Cum hoc ergo propter hoc».
  - La falacia del francotirador.

Falacia «Cum hoc ergo propter hoc»

- «Cum hoc ergo propter hoc» significa «Con ello, luego por ello».
- Que encontremos una relación estadística entre dos variables A y B no significa que una implica a la otra.
- Hay otras razones por las que dos variables pueden estar relacionadas, sin implicación:
  - Fluctuaciones estadísticas **(http://www.tylervigen.com)**
  - Que haya una tercera variable C que implique a las dos. Por ejemplo, la evolución tecnológica en la siguiente imagen.

Falacia «Cum hoc ergo propter hoc»



Falacia «Cum hoc ergo propter hoc»

Falacia del «francotirador»

- Conocida en inglés como «Texas sharpshooter fallacy»
- El nombre viene de imaginar un tirador que dispara aleatoriamente varios tiros a una pared y luego pinta una diana centrada en cada uno de los tiros para presumir de punteria.
- La confusión principal que crea esta falacia es la de tomar los resultados de un estudio exploratorio como si fueran resultados confirmados.
- Los resultados de un estudio exploratorio SIEMPRE han de ser confirmados posteriormente por uno o más estudios confirmatorios.

Falacia del «francotirador»

THE TIMES

UK News

News | Opinion | Business | Money | Sport | Life | Arts | Puzzles | Papers |

The Goals | 15:53:25

Welcome to your preview of The Times

Cocaine floods the playground

By Richard Ford and Sean O'Neill
Published at 12:00AM, March 25 2006

Post a cor
Print
Share via
Facebook
Twitter
Google+

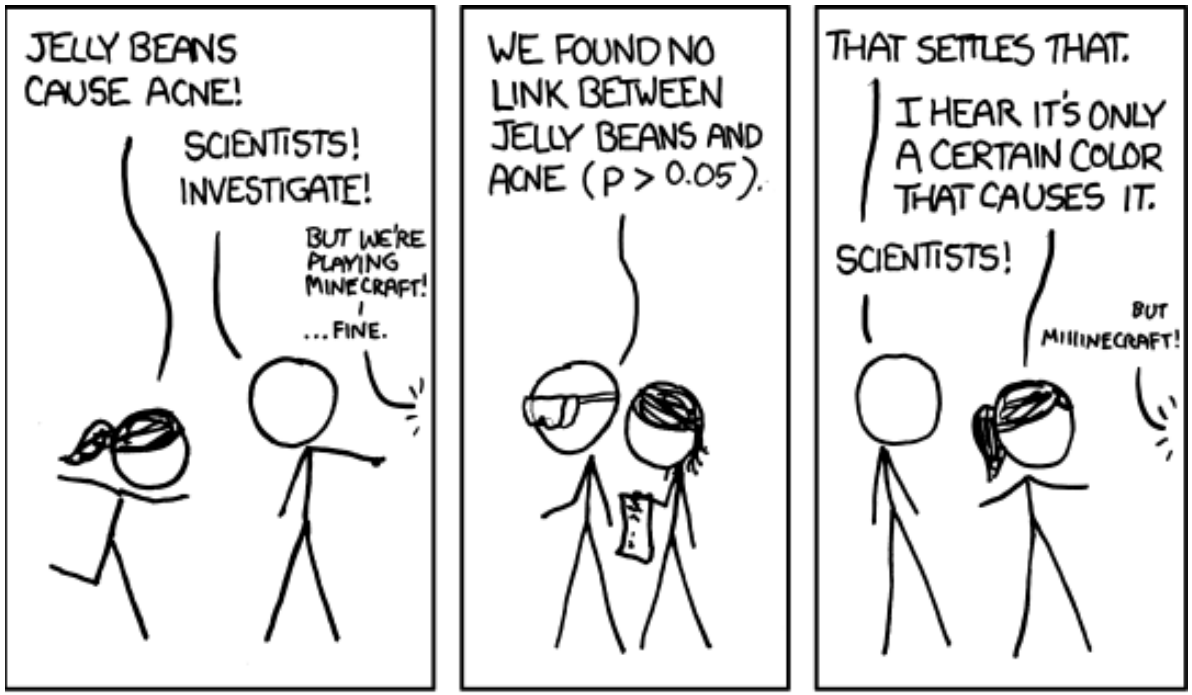
- Use of the addictive drug by children doubles in a year

COCAINE use among children has doubled in a year as the fashionable drug of the middle classes extends its reach from the dinner party to the playground.

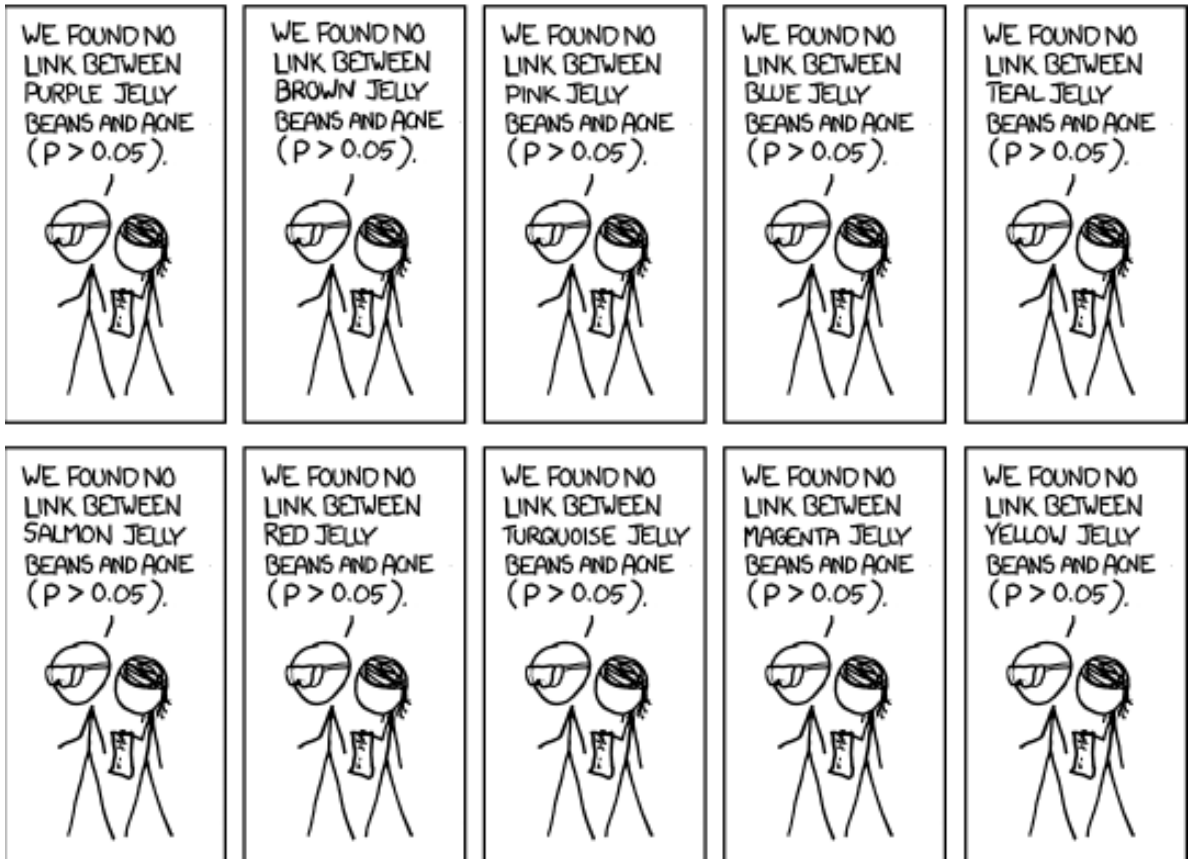
Hundreds of thousands of 11 to 15-year-olds are being offered the Class A drug, which is flooding into the country, according to government figures published yesterday.

An estimated total of 65,000 — one in 50 children aged 11-15 — said they had taken cocaine, which is known by euphemisms such as "zip" and "tickets" by youngsters who are increasingly experimenting

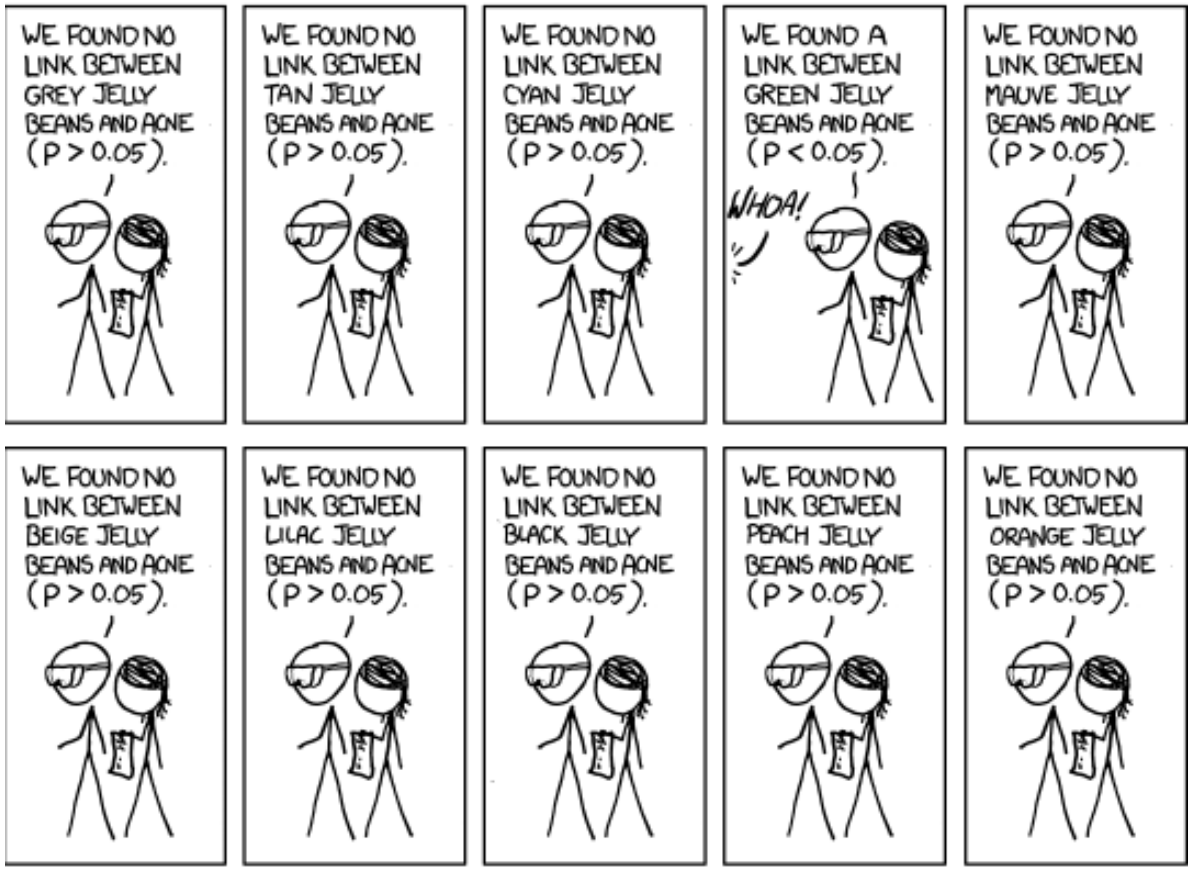
Falacia del «francotirador»



Falacia del «francotirador»



Falacia del «francotirador»



Falacia del «francotirador»