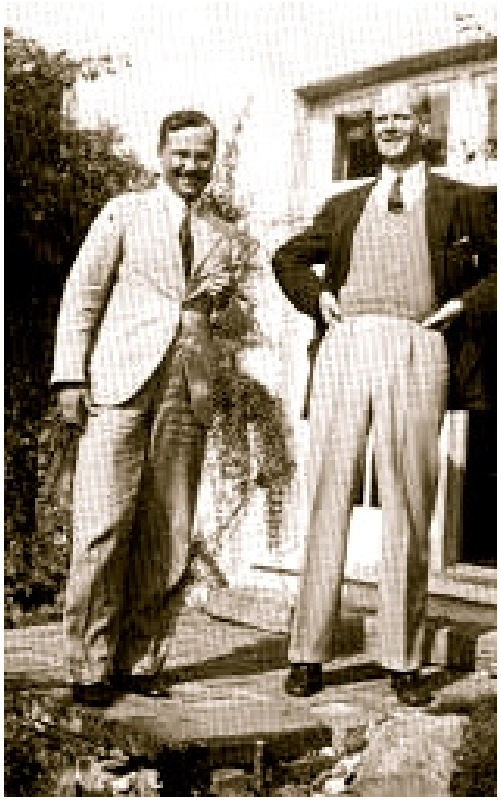
Inferencia clásica o frecuentista



Mike Wiper

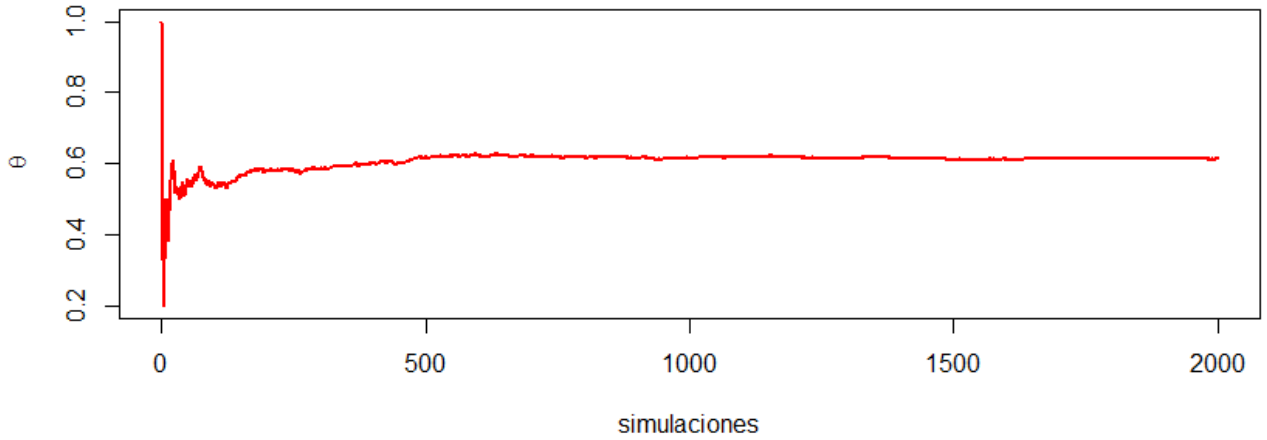
Departamento de Estadística

Universidad Carlos III de Madrid

Objetivo

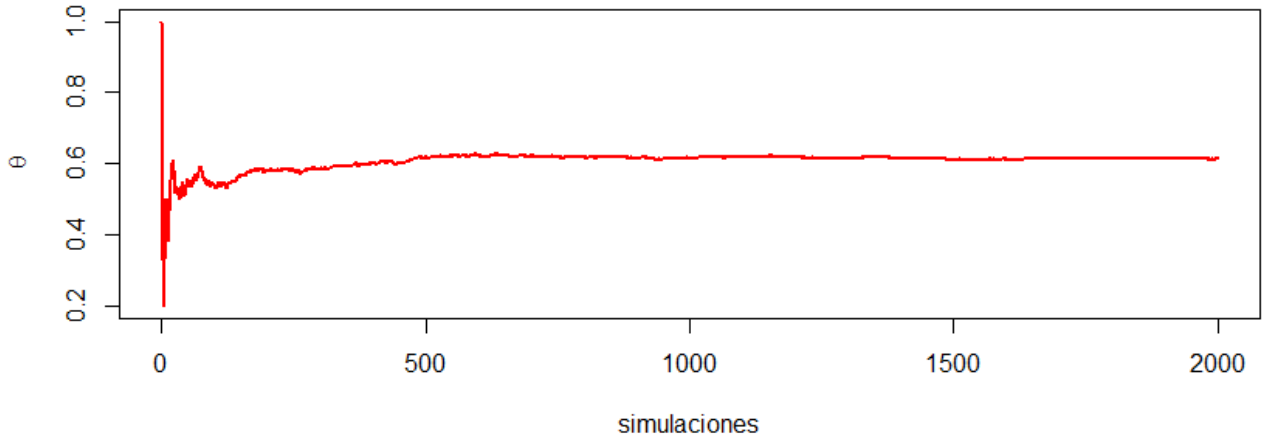
Recordamos como funciona la inferencia frecuentista y ilustrar algunas de sus características.

Probabilidad frecuentista



La inferencia clásica se basa en el uso de la interpretación frecuentista de la probabilidad: en un experimento repetible, la probabilidad de un suceso es el límite de la proporción de ocurrencias del suceso en n repeticiones del experimento cuando $n \to \infty$.

Probabilidad frecuentista



La inferencia clásica se basa en el uso de la interpretación frecuentista de la probabilidad: en un experimento repetible, la probabilidad de un suceso es el límite de la proporción de ocurrencias del suceso en n repeticiones del experimento cuando $n \to \infty$.

¿Tiene sentido usar el concepto para definir la probabilidad de que Real Madrid gane la liga esta temporada?

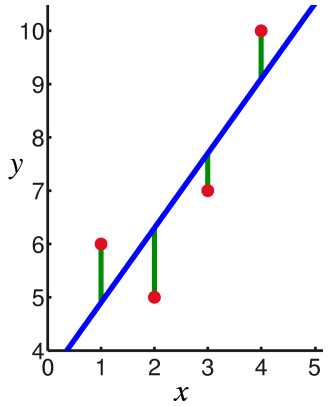
Consecuencias

La probabilidad es un concepto objetivo: la probabilidad de que salga cruz una moneda será la misma para todos.

La probabilidad es un concepto limitado a situaciones de experimentos repetibles.

Parámetros, θ, son desconocidos, pero (típicamente) fijos.

Estimación puntual



Un buen estimador tiene buenas propiedades probabilísticas (frecuentistas):

- Insesgadez,
- Baja varianza,
- Pequeño error cuadrático medio.

Existen varios métodos de seleccionar un estimador, por ejemplo el *método de momentos*, *mínimos cuadrados* o el *estimador máximo verosímil*.

Estimación máximo verosimil

El EMV de un parámetro, θ, es $\hat{\theta}$ tal que:

$$l(\hat{\theta}|\text{data}) = \sup_{\theta \in \Theta} l(\theta|\text{data}).$$

Se sabe que el EMV tiene buenas propiedades *asintóticas*:

- Insesgadez ✓
- Baja varianza ✓
- Pequeño error cuadrático medio. ✓

Sus ventajas no son tan claras en muestras pequeñas.

Ejemplo

Supongamos que decidimos tirar una moneda (con $P(\text{cruz}) = \theta$) 12 veces y observamos 9 cruces.

Entonces, la log verosimilitud es binomial:

BD

$$\log l(\theta|\text{data}) = \log \binom{12}{9} + 9 \log \theta + 3 \log(1-\theta)$$

y derivando,

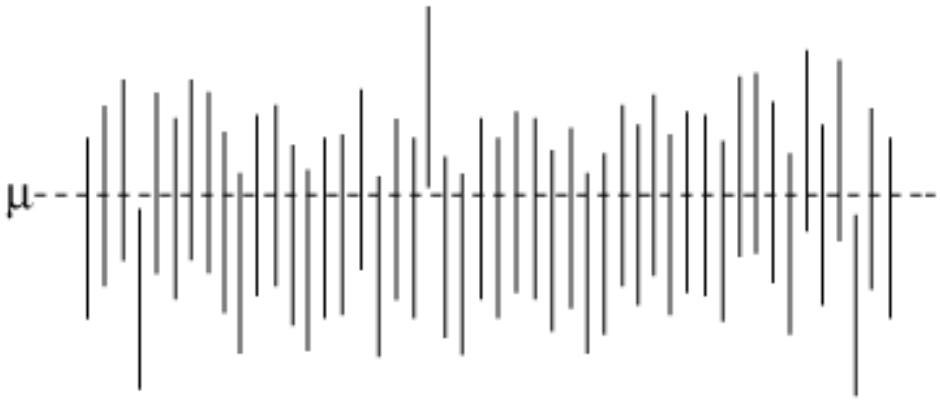
$$\frac{d}{d\theta} \log l(\theta|\text{data}) = \frac{9}{\theta} - \frac{3}{1-\theta}$$

y al máximo,

$$0 = \frac{9}{\hat{\theta}} - \frac{3}{1-\hat{\theta}}$$

que implica que $\hat{\theta} = \frac{9}{12}$.

Estimación por intervalos



Formalmente, un intervalo de $100(1-\alpha)\,\%$ de confianza para un parámetro θ dada una muestra $\mathrm{X} = (X_1, ..., X_n)^T$ es un intervalo aleatorio $(U(\mathrm{X}), V(\mathrm{X}))$ tal que:

$$P(U(\mathrm{X}) < \theta < V(\mathrm{X})) = 1 - \alpha$$

para cualquier valor de θ (y cualquier otro parámetro del modelo).

Ejemplo

Si generamos datos $Y_1, ..., Y_n$, de una distribución normal con media μ y varianza conocidada σ^2, se sabe que la media muestral,

ND

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

y luego,

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

Entonces:

$$\begin{aligned} 0{,}95 &= P\left(-1{,}96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1{,}96\right) \\ &= P\left(\bar{Y} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Dada la muestra $Y_1 = y_1, ..., Y_n = y_n$, el intervalo de 95 % de confianza para μ es:

$$\left(\bar{y} - 1{,}96 \frac{\sigma}{\sqrt{n}}, \; \bar{y} + 1{,}96 \frac{\sigma}{\sqrt{n}} \right).$$

Supongamos que observamos 25 datos de una población con desviación típica conocida $\sigma = 10$ y que la media muestral es $\bar{y} = 4$. Luego el intervalo de confianza es:

$$\left(4 - 1{,}96 \times \frac{10}{\sqrt{25}}, \; 4 + 1{,}96 \times \frac{10}{\sqrt{25}} \right) = (0{,}08, \; 7{,}92).$$

¿Cómo interpretamos este intervalo?

Dada la muestra $Y_1 = y_1, ..., Y_n = y_n$, el intervalo de 95 % de confianza para μ es:

$$\left(\bar{y} - 1{,}96\frac{\sigma}{\sqrt{n}},\ \bar{y} + 1{,}96\frac{\sigma}{\sqrt{n}}\right).$$

Supongamos que observamos 25 datos de una población con desviación típica conocida $\sigma = 10$ y que la media muestral es $\bar{y} = 4$. Luego el intervalo de confianza es:

$$\left(4 - 1{,}96 \times \frac{10}{\sqrt{25}},\ 4 + 1{,}96 \times \frac{10}{\sqrt{25}}\right) = (0{,}08,\ 7{,}92).$$

¿Cómo interpretamos este intervalo?

¿La probabilidad de que contenga μ es 0,95?

Dada la muestra $Y_1 = y_1, ..., Y_n = y_n$, el intervalo de 95 % de confianza para μ es:

$$\left(\bar{y} - 1{,}96\frac{\sigma}{\sqrt{n}},\ \bar{y} + 1{,}96\frac{\sigma}{\sqrt{n}}\right).$$

Supongamos que observamos 25 datos de una población con desviación típica conocida $\sigma = 10$ y que la media muestral es $\bar{y} = 4$. Luego el intervalo de confianza es:

$$\left(4 - 1{,}96 \times \frac{10}{\sqrt{25}},\ 4 + 1{,}96 \times \frac{10}{\sqrt{25}}\right) = (0{,}08,\ 7{,}92).$$

¿Cómo interpretamos este intervalo?

¿La probabilidad de que contenga μ es 0,95? ✗

Contrastes de hipótesis

Un contraste de hipótesis típicamente tiene los siguientes pasos:

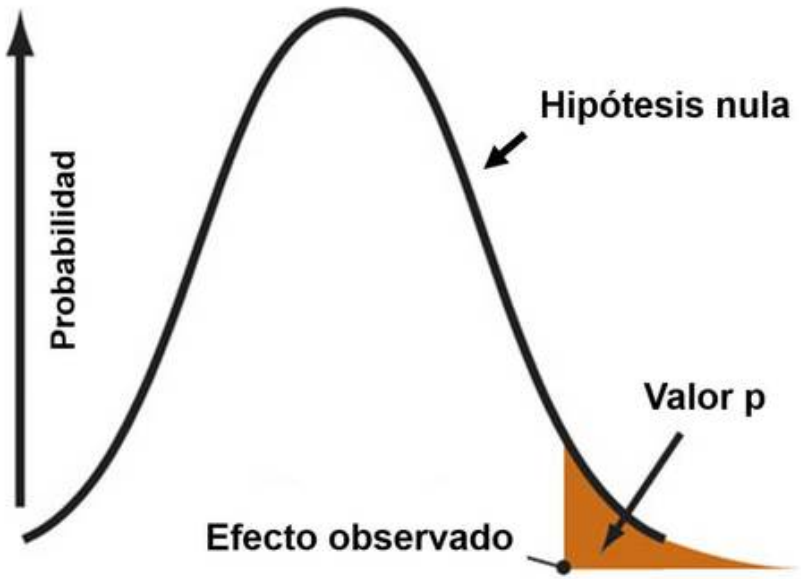
1. Formalizar una hipótesis experimental o alternativa (H_1) de que no se sabe su veracidad.
2. Formalizar la hipótesis opuesta o hipótesis nula (H_0).
3. Derivar un estadístico del contraste.
4. Hallar la distribución del estadístico bajo H_0.
5. Fijar un nivel de significación o error de tipo I:

$$\alpha = P(\text{rechazar } H_0 | H_0 \text{ es verdadera}).$$

6. Calcular la región crítica, es decir el rango de valores del estadístico del contraste donde se rechazaría H_0 a favor de H_1.

Dada la muestra observada, se calcula el valor del estadístico y sí cae dentro de la región crítica, se rechaza H_0.

El p-valor



Como alternativa al último paso, dada la muestra observada, es más común calcular el p-valor, es decir la probabilidad bajo H_0 de observar un valor del estadístico por lo menos tan extremo como el observado.

Si el p-valor es más pequeño que α, se rechaza la hipótesis nula.

A menudo, se interpreta el p-valor como una medida de la fuerza de evidencia en contra de H_0.

Ejemplo

Tenemos una moneda con $P(\text{cruz}) = \theta$ y queremos contrastar $H_0 : \theta = 0{,}5$ frente a $H_1 : \theta > 0{,}5$.

Tiramos la moneda 12 veces y observamos 9 cruces y 3 caras.

Luego, el p-valor es:

$$p = \sum_{x=9}^{12} \binom{12}{x} 0{,}5^x (1 - 0{,}5)^{12-x} = 0{,}073$$

y no rechazamos la hipótesis nula a un nivel de significación de 5 %.

Un resultado paradójico

Supongamos que en lugar de decidir de antemano tirar la moneda 12 veces, habíamos dicho que tiraríamos la moneda hasta observar la tercera cara.

Un resultado paradójico

Supongamos que en lugar de decidir de antemano tirar la moneda 12 veces, habíamos dicho que tiraríamos la moneda hasta observar la tercera cara.

Ya el diseño del experimento es binomial negativa en lugar de binomial.

Un resultado paradójico

Supongamos que en lugar de decidir de antemano tirar la moneda 12 veces, habíamos dicho que tiraríamos la moneda hasta observar la tercera cara.

Ya el diseño del experimento es binomial negativa en lugar de binomial.

Si la tercera cara ocurre en la duodécima tirada, todavía hemos visto 9 cruces y 3 caras y el EMV de θ sigue siendo $\hat{\theta} = \frac{9}{12}$.

Un resultado paradójico

Supongamos que en lugar de decidir de antemano tirar la moneda 12 veces, habíamos dicho que tiraríamos la moneda hasta observar la tercera cara.

Ya el diseño del experimento es binomial negativa en lugar de binomial.

Si la tercera cara ocurre en la duodécima tirada, todavía hemos visto 9 cruces y 3 caras y el EMV de θ sigue siendo $\hat{\theta} = \frac{9}{12}$.

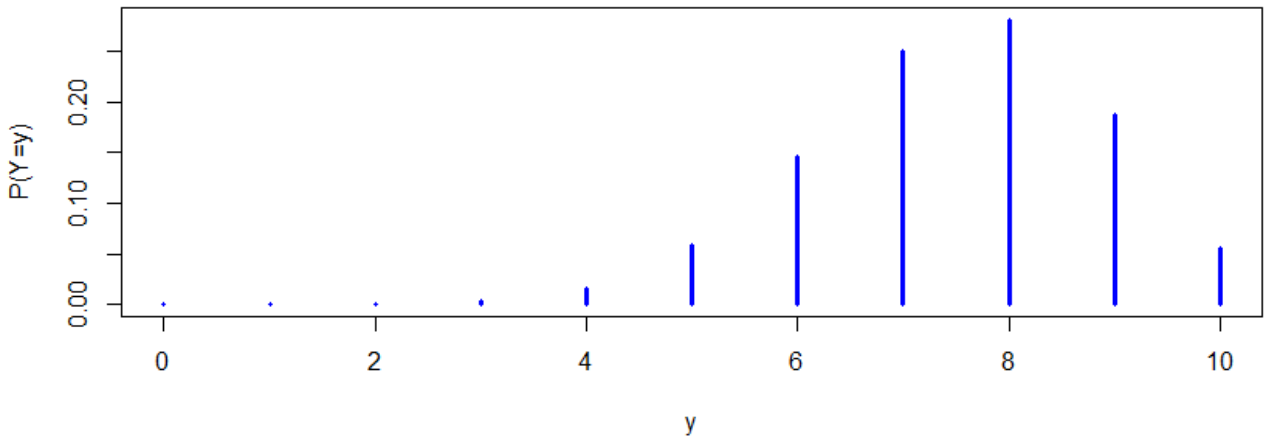
No obstante, ahora el p-valor es

$$p = \sum_{y=9}^{\infty} \binom{y+3-1}{y} (1-\theta)^3 \theta^9 = 0{,}033$$

y sí rechazamos la hipótesis nula.

¿Os parece lógico?

Predicción



Típicamente, para hacer predicción en un procedimiento clásico, se utiliza un método "plug in".

Por ejemplo, si queremos predecir $Y =$ el número de cruces en 10 tiradas más de la moneda, estimariamos que

$$Y \sim \text{Binomial}(10, 0{,}75).$$

Comparación de modelos

Para comparar varios modelos, se puede utilizar, por ejemplo un criterio basado en penalizando la verosimilitud de acuerdo con el número de parámetros.

Por ejemplo, el AIC para un modelo $\mathcal{M}$ es:

$$AIC = -2 \log l\left(\hat{\theta}_{\mathcal{M}} | \text{data}, \mathcal{M}\right) + 2k$$

dónde k es el número de parámetros en el modelo.

El modelo seleccionado sería el que minimice el AIC.

Ejemplo

Volviendo al ejemplo, supongamos que queremos comparar el modelo con $p = 0{,}5$, $(\mathcal{M}_0)$ con el modelo completo $(\mathcal{M}_1)$: $Y \sim \text{Binomial}(12, p)$ para cualquier p.

El AIC para el modelo $\mathcal{M}_0$ es

$$AIC_0 = -2 \log \left\{ \binom{12}{9} 0{,}5^{12} \right\} + 2 \times 0 = 5{,}85$$

y el AIC para el modelo general es

$$AIC_1 = -2 \log \left\{ \binom{12}{9} 0{,}75^9 0{,}25^3 \right\} + 2 \times 1 = 4{,}71.$$

Luego el modelo preferido por el AIC es el modelo completo.

Bondad de ajuste

Para ver si los datos se ajustan a un modelo $\mathcal{M}$, se puede utilizar un contraste de hipótesis de la hipótesis nula:

$$H_0 : \text{los datos provienen del modelo } \mathcal{M}.$$

Ejemplos típicos son el contraste χ^2 o el contraste *Kolmogorov-Smirnov*.

Ejemplo

Una posibilidad es hacer un contraste de razón de verosimilitudes.

Asintoticamente, la distribución de dos veces la diferencia entre los log-verosimilitudes del modelo saturado o completo y del modelo que se quiere contrastar es χ^2_k donde k es la diferencia entre el número de parámetros de los dos modelos.

▸ χ^2

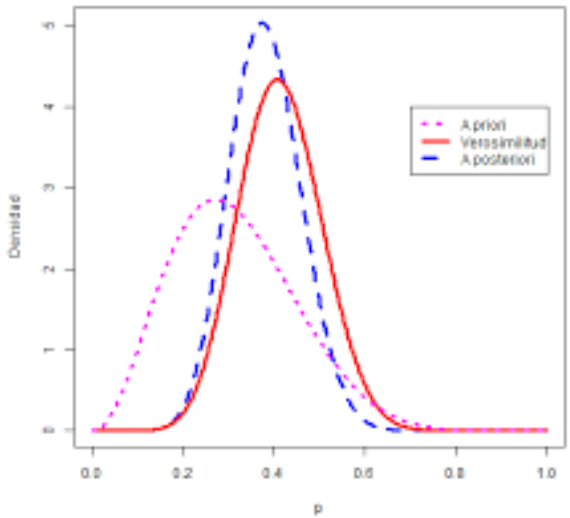
En nuestro caso, la diferencia en log-verosimilitudes es

$$2\left(-1{,}35-\left(-2{,}92\right)\right)=3{,}14$$

y el punto crítico de la distribución χ^2_1 es 3.84 y luego no hay suficiente evidencia para decir que el modelo con $p=0{,}5$ no se ajusta a los datos.

Resumen y siguiente sesión

En esta clase, hemos resumido los puntos más característicos de la inferencia clásica o frecuentista.



En la siguiente sesión, introduciremos las ideas básicas de la inferencia bayesiana.

Apéndice: la distribución binomial

Una variable discreta, X, tiene una distribución binomial con parámetros $n \in \{1, 2, ...\}$ y $0 < p < 1$ si:

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \quad \text{para } x = 0, 1, 2, ..., n.$$

Se tiene $E[X] = np$ y $V[X] = np(1-p)$.

Apéndice: la distribución normal

Una variable continua, Y, tiene una distribución con parámetros $\mu \in \mathbb{R}$ y $\sigma^2 > 0$ si:

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \quad \text{para } y \in \mathbb{R}.$$

Se tiene $E[Y] = \mu$ y $V[Y] = \sigma^2$.

La distribución normal con $\mu = 0$ y $\sigma^2 = 1$ es la distribución normal estandár.

Apéndice: la distribución ji-cuadrado

Una variable continua, Y, tiene una distribución ji-cuadrado con k grados de libertad (donde $k \in \{1, 2, ...\}$) si:

$$f_Y(y) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} \exp\left(-\frac{y}{2}\right) \quad \text{para } y > 0.$$

Se tiene $E[Y] = k$ y $V[Y] = 2k$.

La distribución χ^2_k es igual a la distribución gamma$\left(\frac{k}{2}, \frac{1}{2}\right)$.