Métodos Numéricos y Simulaciones en Astrofísica

Parte 4: Estadística para Astronomía

Estadística para Astronomía

- Estimación de errores.
- Estadísticos (media, moda, mediana, etc.)
- Inferencia estadística.
- Método científico.

Pero en astronomía:

- No se puede controlar los experimentos.
- No se pueden repetir los experimentos.
- Usualmente las muestras son pequeñas.
- Tendencia a verificar hipótesis usando los mismos datos con los que se formuló la misma.

Estadística para Astronomía

El análisis estadístico clásico indica que:

1. Debemos formular una hipótesis
2. Recolectar datos mediante experimentos
3. Construir un estadístico

Para obtener decisiones a partir del estadístico debemos conocer como se distribuye la muestra

En astronomía se deben usar

probabilidades y métodos no-paramétricos

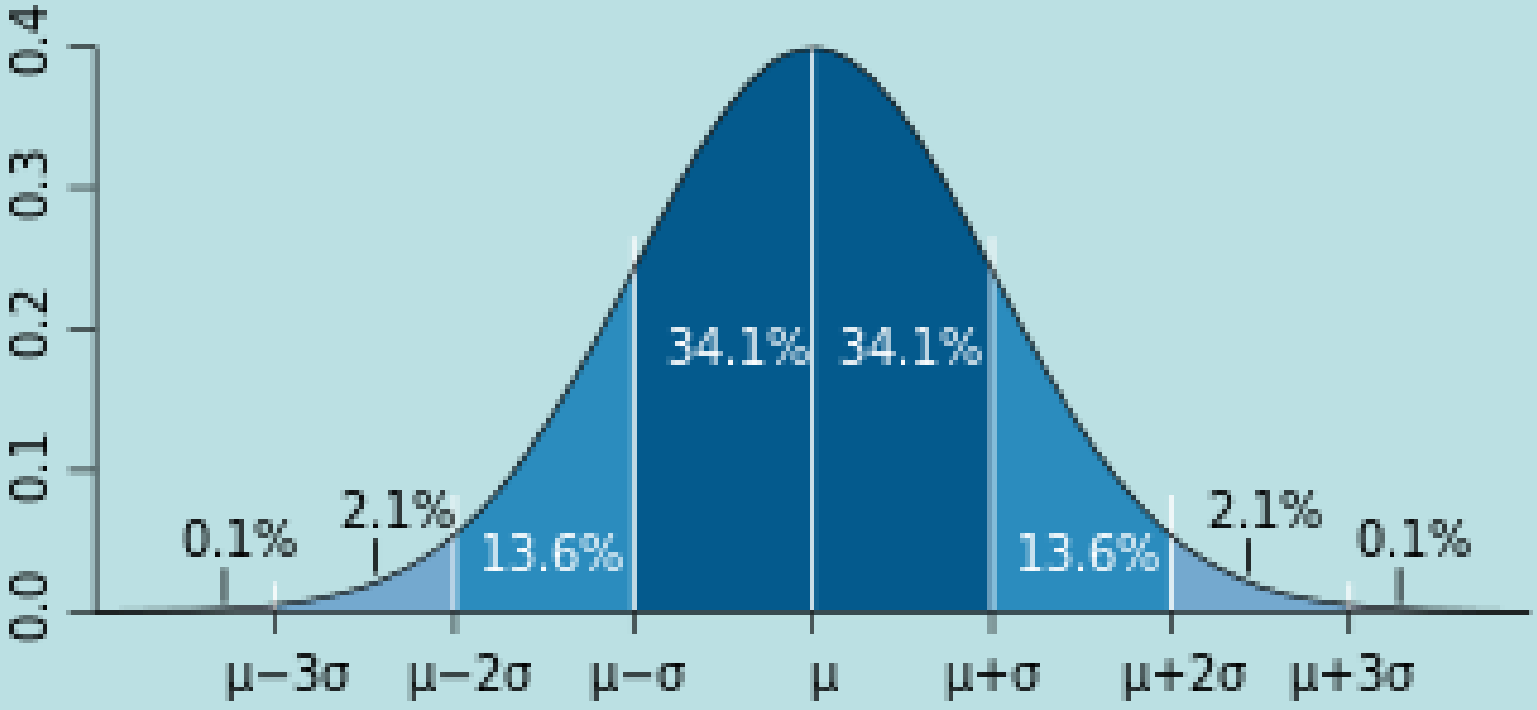
Estadística para Astronomía

Qué significado tiene la notación:

$$|P_{min}| = 1.27 \pm 0.20\,\%$$

$$\alpha_{min} = 8.6 \pm 1.2^{\circ}$$

Estadística para Astronomía



Probabilidad del 68.2% que la cantidad se encuentre a menos de 1σ del valor asignado (95.4% a 2σ y 99.6% a 3σ)

Teorema del límite central

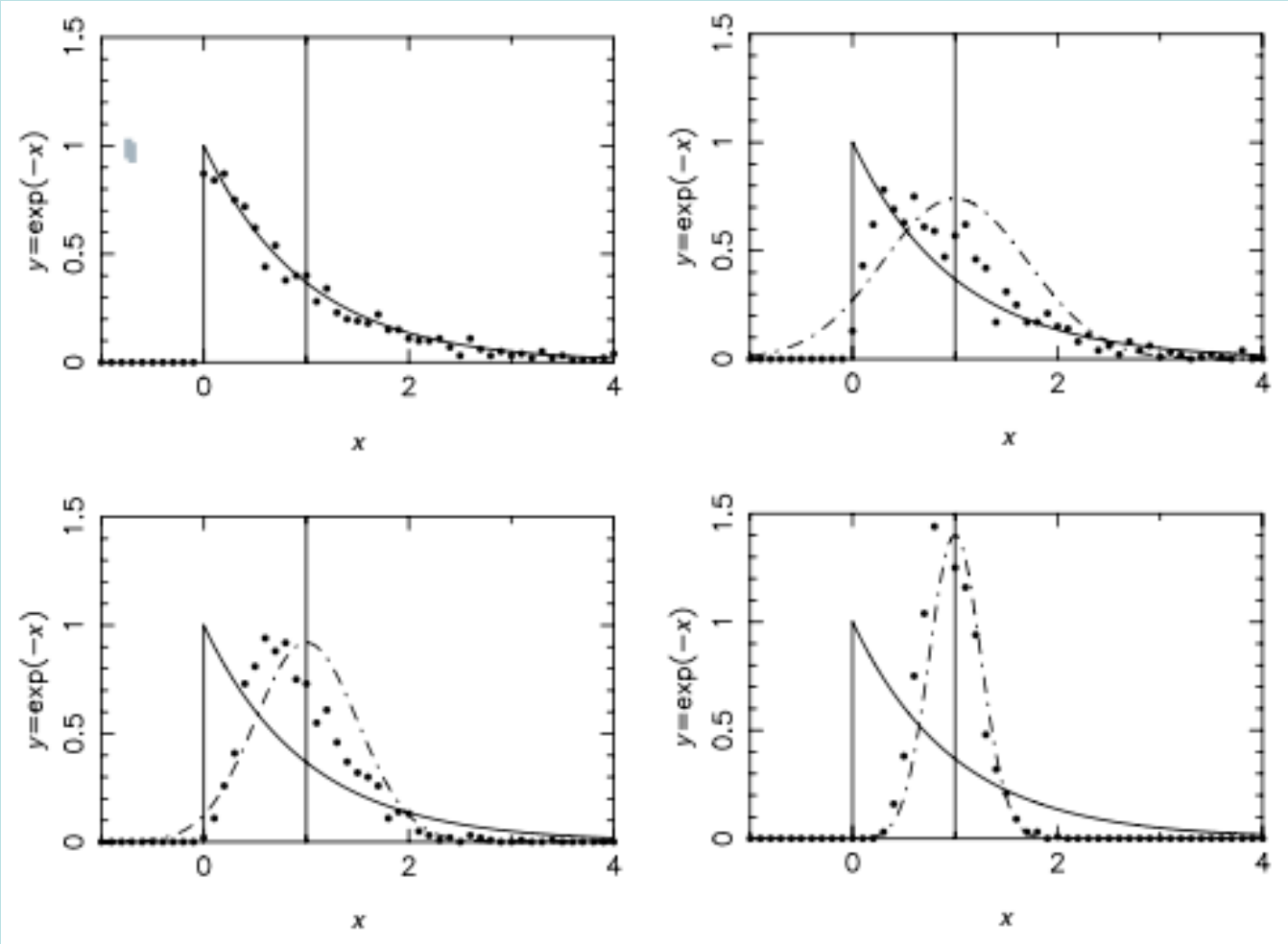
Teorema (del límite central): Sea $X_1, X_2, \ldots, X_n$ un conjunto de variables aleatoria, independientes e idénticamente distribuidas de una distribución con media **μ** y varianza $\sigma^2 \neq 0$.

Entonces, si **n** es suficientemente grande, la variable aleatoria

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$.

Teorema del límite central



200 valores de una exponencial truncada promediados de a 1, 2, 4 y 16 valores

Probabilidades

La probabilidad es una formalización numérica de nuestro grado de creencia.

Axiomas de Kolmogorov:

1. Cualquier evento al azar A tiene prob(A) en el rango [0,1].
2. El evento seguro tiene prob(A)=1.
3. Si A y B son eventos mutuamente excluyentes, prob (A and B)=prob(A)+prob(B).

Probabilidades

- Si dos eventos A y B son independientes, entonces prob (A and B)=prob(A) * prob(B).

- Si dos eventos A y B no son independientes entonces prob(A|B) = prob(A and B) / prob(B) (probabilidad condicional).

Probabilidades

Distribution	Density function	Mean	Variance
Uniform	$f(x;a,b) = 1/(b-a)\; a < x < b$ $= 0, \quad x < a, x > b$	$(a+b)/2$	$(b-a)/12$
Binomial	$f(x;p,q) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$	np	npq
Poisson	$f(x;\mu) = \frac{e^{-\mu}\mu^x}{x!}$	μ	μ
Normal (Gaussian)	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2
Chi-square	$f(\chi^2;\nu) = \frac{\chi^{2(\nu/2-1)}}{2^{\nu/2}\Gamma(\nu/2)} \exp(-\frac{1}{2}\chi^2)$	ν	2ν
Student t	$f(t;\nu) = \Gamma[(\nu+1)/2] \frac{(1+t^2/\nu)^{-[(\nu+1)/2]}}{\sqrt{\pi\nu}\Gamma(\nu/2)}$	0	$\nu/(\nu-2)$ (for $\nu > 2$)

Probabilidades:
Método Frecuentista

- Los estadísticos son combinaciones de datos que no dependen de parámetros desconocidos.
- Se asume que los estadísticos guardan alguna relación con los parámetros de la distribución de donde se tomaron los datos.
- Podemos calcular su distribución si repetimos el experimento muchas veces.
- Conocemos la probabilidad de que el valor real se encuentre dentro de un cierto rango del valor encontrado.

Probabilidades:
Método Frecuentista

- La probabilidad de cierto evento es el cociente entre el número de casos favorables y el número total de eventos.
- Se asigna igual probabilidad a los eventos a menos que se disponga de información que los distinga (Principio de Indiferencia).
- Es usual definir probabilidad mediante frecuencias para un gran número de casos, pero al seleccionar la muestra se deben conocer siempre los casos de igual probabilidad, lo que lleva a una definición circular.

Probabilidades:
Método Bayesiano

- No se realiza un paso estadístico.
- El razonamiento es inverso a la aproximación frecuentista.
- Los datos son únicos y conocidos.
- Los parámetros son desconocidos y se le asigna una probabilidad.
- Sin usar la estadística, se calcula la probabilidad de varios valores para el parámetro buscado a partir de los datos obtenidos.

Probabilidades:
Método Bayesiano

Igualando prob(A and B) con prob(B and A) se obtiene el Teorema de Bayes:

$$prob(B|A) = prob(A|B)\ prob(B) / prob(A)$$

El teorema es particularmente útil como regla de inducción: los datos (A) se producen dado el estado conocido (B).

Probabilidades:
Método Bayesiano

- prob(B|A) es la probabilidad posterior.
- prob(A|B) es la esperanza que se obtiene de la experiencia.
- prob(B) es la probabilidad a priori.
- prob(A) es un factor de renormalización.

Probabilidades

Ejemplo frecuentista:

En una urna tenemos N=6 bolas rojas y M=4 bolas azules. Si en T=5 intentos sacamos con reposición R=3 bolas rojas y A=2 azules, cuál es la probabilidad de este resultado?

Es una distribución binomial, entonces la probabilidad es:

$$\binom{T}{R}\left(\frac{N}{N+M}\right)^{R}\left(\frac{M}{N+M}\right)^{T-R}.$$

prob=0.346

Probabilidades

Ejemplo bayesiano:

En una urna tenemos N bolas rojas y M bolas azules, con N+M=10. Si en T=5 intentos sacamos con reposición R=3 bolas rojas y A=2 azules, cuántas bolas rojas hay en la urna?

Es una distribución binomial, entonces la probabilidad de obtener esos datos dado el modelo, prob(A|B), es:

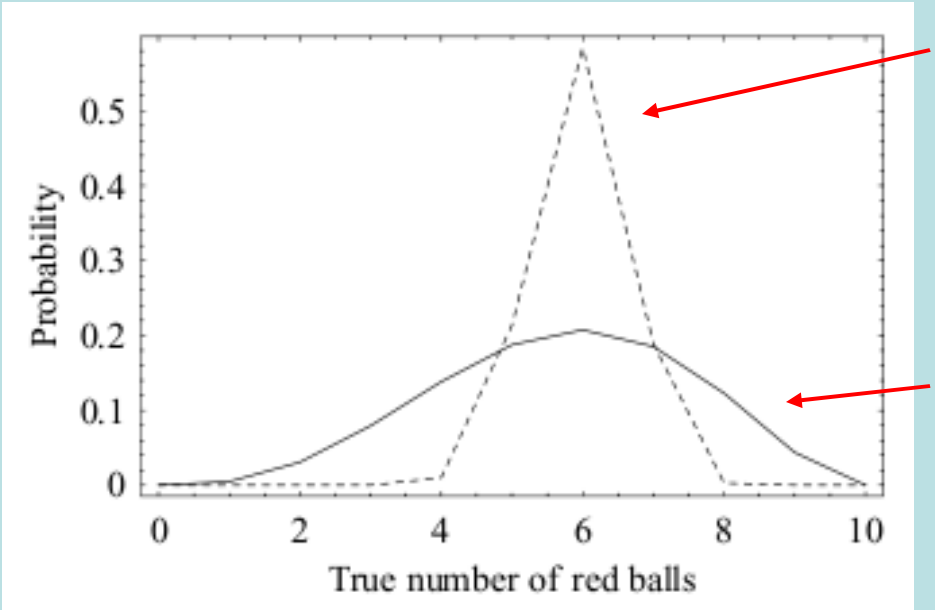
$$\binom{T}{R}\left(\frac{N}{N+M}\right)^{R}\left(\frac{M}{N+M}\right)^{T-R}.$$

Probabilidades

prob(B), es la probabilidad de obtener N. Como no hay más información tomamos N uniforme entre 1 y N+M=10.

prob(A), se obtiene normalizando prob(B|A).

Entonces, la probabilidad posterior, prob(B|A) (el número de rojas a partir de los intentos), es:



Probabilidades

Otro ejemplo bayesiano:

Detectamos una fuente con flujos f los cuales corresponderían a un flujo verdadero S que se distribuye en forma normal con varianza 1. Si los flujos observados son [2., 1.3, 3., 1.5, 2., 1.8], cuánto vale S?

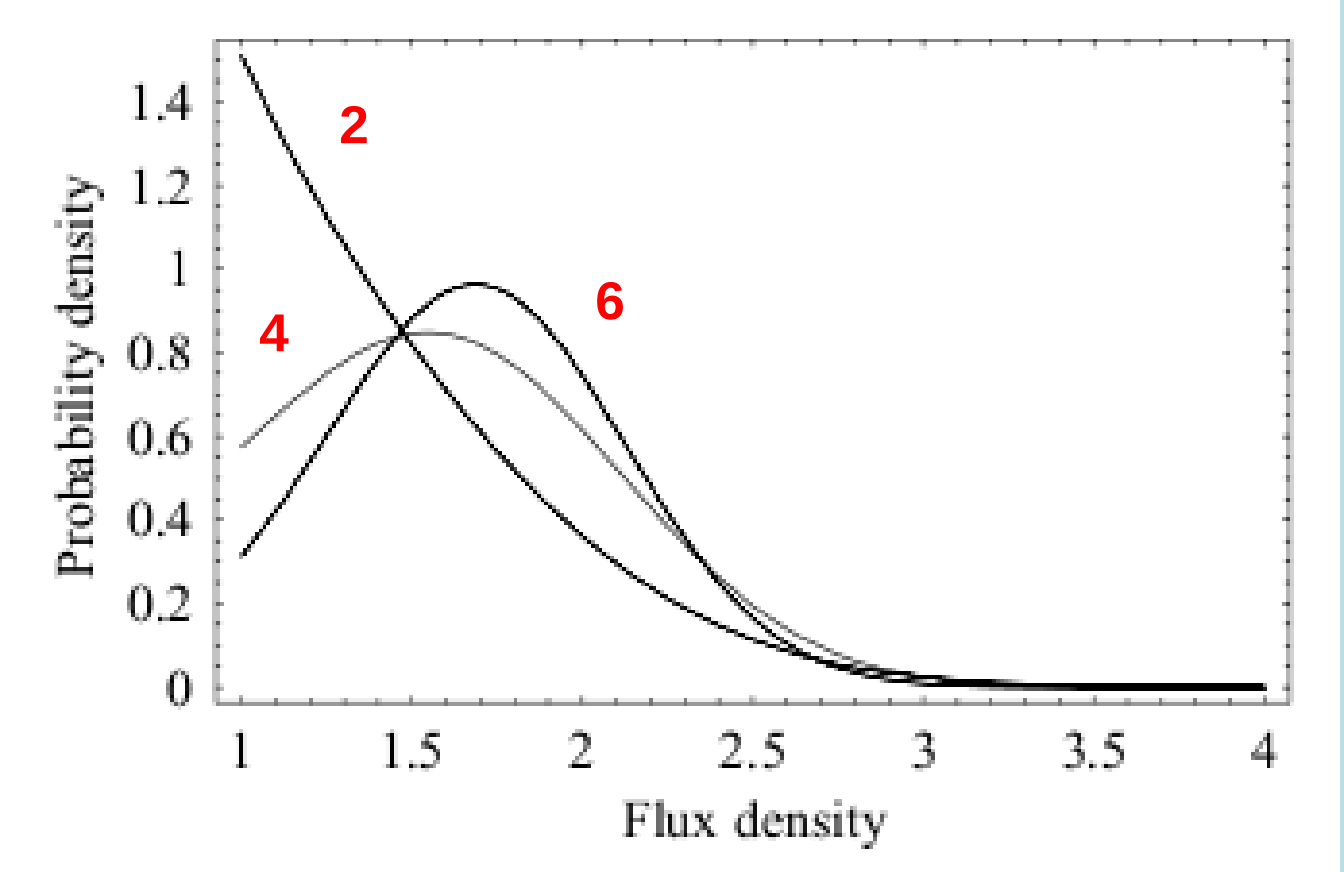
Probabilidades

- La esperanza es normal:

$$\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(f_i - S)^2\right]$$

- La probabilidad a priori es desconocida. Usemos S^(-5/2) suponiendo que hay más objetos de flujo bajo que alto.
- Calculemos para 2, 4 y 6 mediciones.

Probabilidades



La elección de la probabilidad a priori afecta el resultado para pocas mediciones

Prueba de Hipótesis

1. Se definen dos hipótesis excluyentes: H_0, la hipótesis nula formulada para rechazarla, y H_1, la hipótesis alternativa.
2. Se fija a priori un nivel de significación α y se elije una prueba que: a) aproxima las condiciones fijadas y, b) encuentra la distribución muestral y la zona de rechazo.
3. Se ejecuta la prueba y H_0 se rechaza si el estadístico obtenido es $< \alpha$.
4. La prueba rechaza H_0 , pero no acepta H_1.

Prueba de Hipótesis

- Las pruebas de hipótesis que se basan en propiedades de la distribución de donde se sacan los datos se denominan pruebas paramétricas.
- Existen pruebas paramétricas clásicas (frecuentistas) y bayesianas.
- Las pruebas de hipótesis que no se basan en propiedades de la distribución de donde se sacan los datos se denominan pruebas no paramétricas.

Prueba de Hipótesis

	Parametric	Non-parametric
Bayesian testing	Model known. Data gathering and uncertainty understood.	Such tests do not exist.
Classical testing	Model known. Underlying distribution of data known. Large enough numbers. Data on ordinal or interval scales.	Small numbers. Unknown model. Unknown underlying distributions or errors. Data on nominal or categorical scales.

Pruebas no-paramétricas: simple

Test	Applicability†	$N < 10$?	Comment
Binomial test	Goodness-of-fit (N)	Yes	Appropriate for two-category (dichotomous) data; do *not* dichotomize continuous data.
*Chi-square test	Goodness-of-fit (N)	No	For testing categorized, pre-binned, or classified data; choose categories with expected frequencies 6–10.
*Kolmogorov–Smirnov one-sample test	Goodness-of-fit (O)	Yes	The most powerful test for data from a continuous distribution; may always be more efficient than the chi-square test.
*One-sample runs test	Randomness of event sequences (O)	Yes	Does not estimate differences between groups.
Change-point test	Change in the distribution of an event sequence (O)	Yes	Robust with regard to changes in distributional form; efficient.

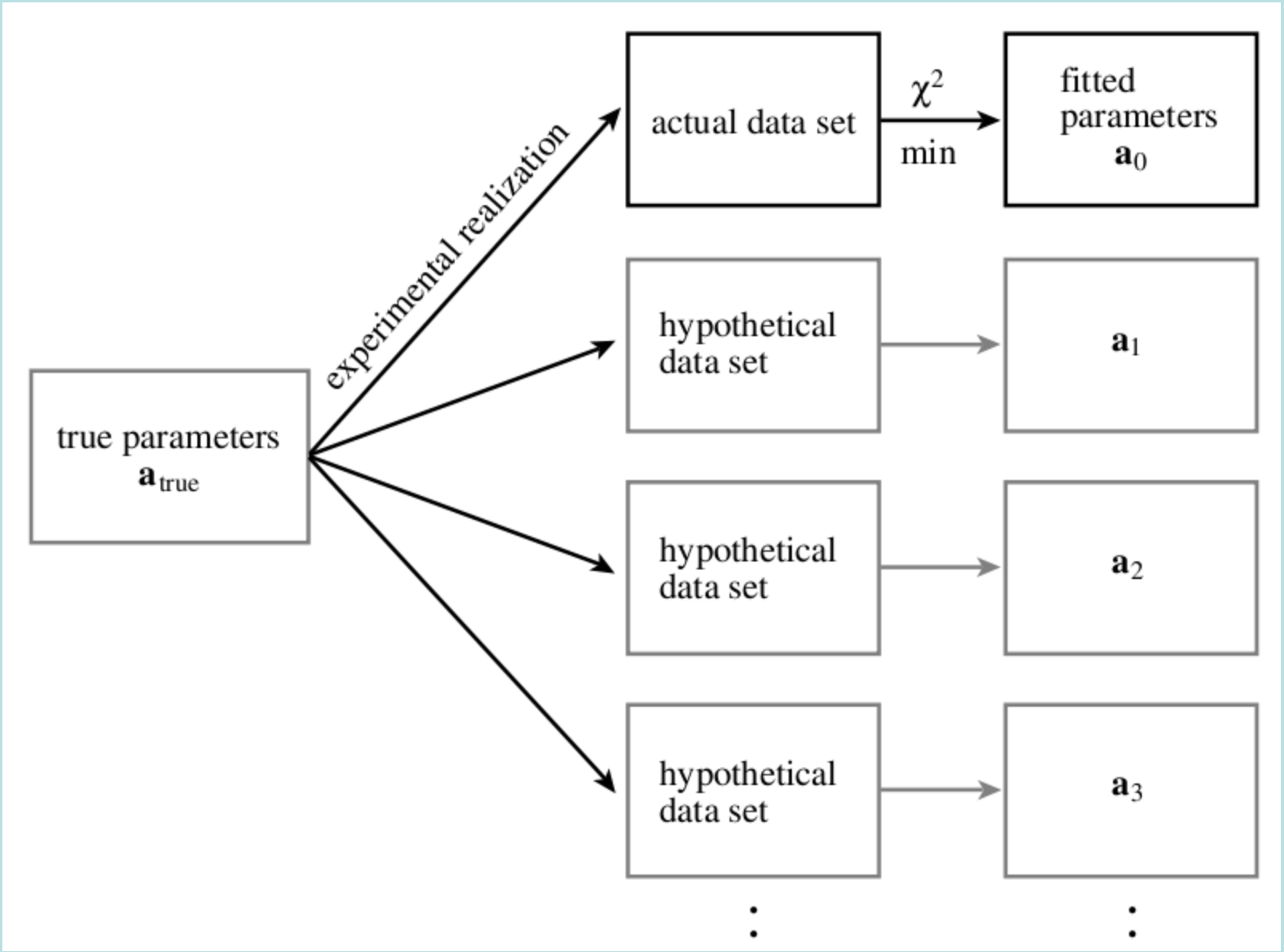
Pruebas no-paramétricas: doble

Test	Applicability†	$N < 10$?	Comment
*Fisher exact test for 2×2 tables	Difference (N)	Yes	The most powerful test for dichotomous data.
*Chi-square test for $r \times 2$ tables	Difference (N)	No	Best for pre-binned, classified, or categorized data.
Median test	Location (O)	Yes	Best for small numbers; efficiency *decreases* with N.
*U (Wilcoxon–Mann–Whitney) test	Location (O)	Yes	One of the most efficient non-parametric tests.
Robust rank-order test	Location (O)	Yes	Efficiency similar to U test.
*Kolmogorov–Smirnov two-sample test	Two-tailed: Difference One-tailed: Location (O)	Yes	The most powerful test for data from a continuous distribution.
Siegel–Tukey test for scale differences	Dispersion (O)	Yes	The medians must be the same (or known) for both distributions. Low efficiency.

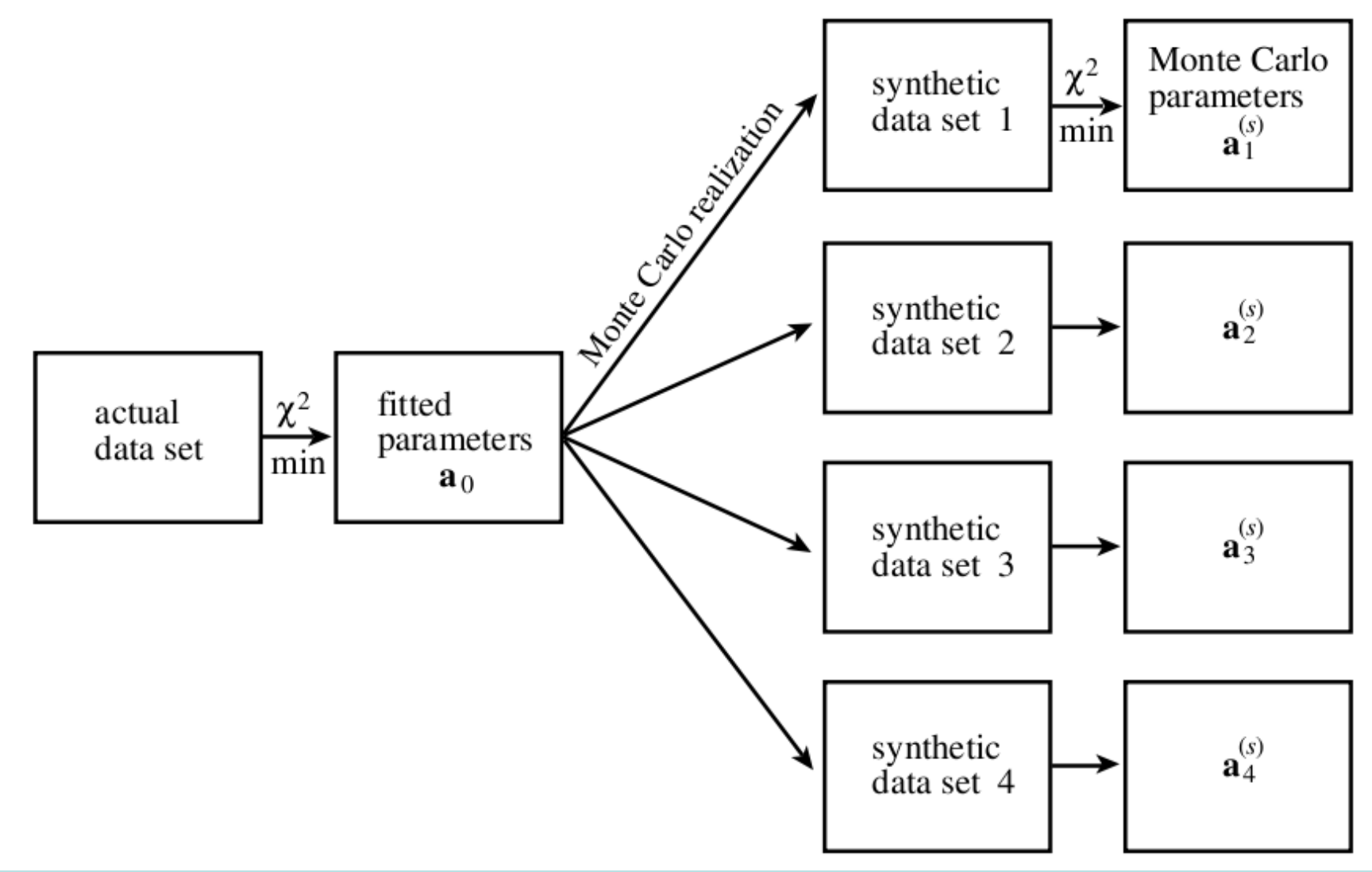
Simulaciones por Monte Carlo

- Los datos observacionales son una muestra parcial de la población real de distribución desconocida.
- Usualmente, se pretende ajustar estos datos observacionales a un modelo dado por ciertos parámetros **a**.
- No se sabe como afectan los errores en los datos la determinación de **a**.
- Entonces, cuáles son los errores en los parámetros del modelo?.

Simulaciones por Monte Carlo



Simulaciones por Monte Carlo



Simulaciones por Monte Carlo

- Bootstrap:

1. Encontrar una distribución para la muestra. Encontrar M valores al azar (sobre N).
2. Reemplazar en la muestra los M valores Recalcular **a** (los designamos **a'**).
3. Repetir el proceso muchas veces.
4. Calcular los valores medios y desviación standard de **a'**.

Simulaciones por Monte Carlo

- Jackknife:

1. Encontrar una distribución para la muestra.
2. Eliminar un valor de la muestra. La muestra ahora tiene N-1 elementos.
3. Recalcular N veces **a** (los designamos $\mathbf{a}_j$).
4. Se encuentran los pseudo-parámetros dados por $\hat{\mathbf{a}} = N\,\mathbf{a} - (N-1)\,\mathbf{a}_j$.
5. Calcular los valores medios y desviación standard de $\hat{\mathbf{a}}$.