Calibrando modelos a datos
Fitting models to data

Billy Ernst
Departamento de Oceanografía
Universidad de Concepción
Chile

Los componentes

1. Modelo de dinámica poblacional
2. Condiciones iniciales
3. Observaciones (datos)
4. Modelo de observación
5. Función objetivo

Los componentes

$X_{t+1} = f(X_t)$	Modelo de dinámica poblacional
$X_1 = I$	Condiciones iniciales
$\{O_1, \dots O_T\}$	Datos
$\tilde{O}_t = g(X_t)$	Modelo de Observación
$h(O, \tilde{O})$	Función Objetivo

Modelo de producción de Schaeffer

$$B_{t+1} = f(B_t) = B_t + rB_t(1 - B_t / K) - C_t$$

Modelo de dinámica poblacional

$$B_1 = K$$

Condición inicial: Stock no explotado

$$\{CPUE_1, \ldots CPUE_T\}$$

Data: Indice relativo de abundancia (CPUE)

$$\tilde{O}_t = g(B_t) = qB_t$$

Modelo de observación: CPUE proporcional a la abundancia

$$h(CPUE, \tilde{O}) = \sum_t (CPUE_t - \tilde{O}_t)^2$$

Función objetivo: Mínimos cuadrados

Nota: Captura es un forzante, medido sin error

Precaución: Modelo de Schaefer → MSY ocurre en K/2, supuesto fuerte

Error de Proceso

$$X_{t+1} = f\left(X_t\right)$$

Modelo determinístico

$$X_{t+1} = f\left(X_t\right) + \varepsilon_t$$

Modelo estocástico

$$B_{t+1} = \left[B_t + rB_t\left(1 - B_t / K\right)\right]e^{\varepsilon_t} - C_t$$

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

Ejemplo de Schaeffer

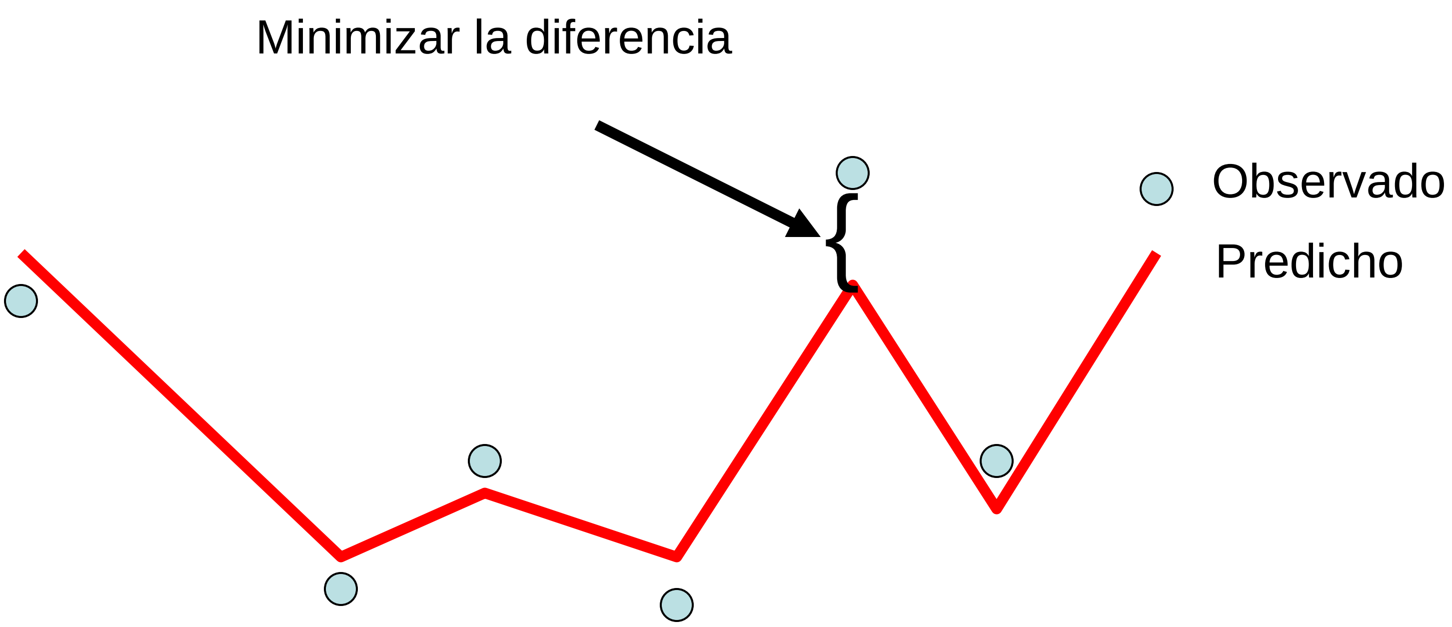
Otros componentes de la evaluación de stock

1. Métodos de estimación
   - Momentos
   - Mínimos cuadrados
   - Máxima verosimilitud
   - Bayesiano
2. Estimación de incertidumbre
   - Método delta
   - Bootstrap
   - Perfil de verosimilitud (likelihood)
   - Distribución a posteriori (Bayesiano)
3. Parámetros de manejo
   - MSY
   - Nivel de depleción
4. Análisis de decisión
   - Proyecciones

Estimando los parámetros del modelo

- Error de observación
- Verosimilitud (Likelihood)
- Error de proceso

Ajustando el modelo a los datos



Como medir la diferencia entre lo observado y lo esperado

- Mínimos cuadrados
- Valor absoluto de la diferencia
- Verosimilitud
- Separación entre error de observación y de proceso

Probabilidad y verosimilitud

Fuentes de Incertidumbre (Francis and Shotten)

- **Error de proceso** (variable en la dinámica espacial y temporal del recurso y la pesquería)
- **Error de observación** (error de muestreo y de medición)
- **Error de modelo** (capacidad del modelo de capturar la dinámica fundamental del sistema)
- **Incertidumbre en la estructura del error** (error en la especificación de las funciones de verosimilitud)
- **Error de estimación** (error en la especificación de los parámetros del modelo)
- **Error de implementación** (efectos de las acciones de manejo difieren de lo originalmente establecido)

Error de proceso

- Emerge de la variabilidad natural. El ejemplo más común de incertidumbre de proceso es la variación en el reclutamiento por causas ambientales
- En modelos de captura a la edad el error de proceso se asocia casi exclusivamente a la varibilidad del reclutamiento
- En el modelo de la merluza común se incorpora cambios temporales en la mortalidad natural

Error de observación

- Emerge del error de muestreo y medición
- En los modelos de captura a la edad se incorpora error de medición para cada pieza de información

Incertidumbre de modelo

- Emerge cuando existe una falta de comprensión sobre la dinámica subyacente del sistema de estudio
- Todos los modelos son simplificaciones y siempre debemos que terminar con una estructura de modelo específica
- Pueden existir diferencias importantes entre el modelo y la realidad
- Habitualmente no es considerada en las evaluaciones de stock, de tal forma que termina siendo asimilada dentro de error de proceso y observación.

Incerteza en la estructura del error

- Se genera como parte de nuestra incapacidad de especificar correctamente el modelo del error
- Es nuestra función de verosimilitud normal o lognormal? Multinomial o Normal robusta para proporciones?
- Absorvida en el error de proceso y observación

Error de estimación

- Emerge de la estimación de parámetros. En parte es función de las cuatro previas fuentes de incertidumbre.
- Se refleja en la incertidumbre asociada a los parámetros estimados
- Reparametrizaciones

Error de implementación

- Es la implicancia de fallar en la implementación acciones de manejo acordadas. En la última década se ha reconocido su importancia en el análisis de decisión en pesquerías.
- Se considera a menudo en MSE

Estimación de parámetros

- El estudio de sistemas biológicos requiere a menudo del **uso de modelos** y de la **obtención de datos** (muestreos y/o experimentos) apropiados para parametrizar dichos modelos. Esta parametrización requiere de un *procedimiento estadístico formal* y *objetivo* para estimar los parámetros del modelo. Si bien existen varios métodos estadísticos que permiten alcanzar este objetivo, son tres los de mayor popularidad:

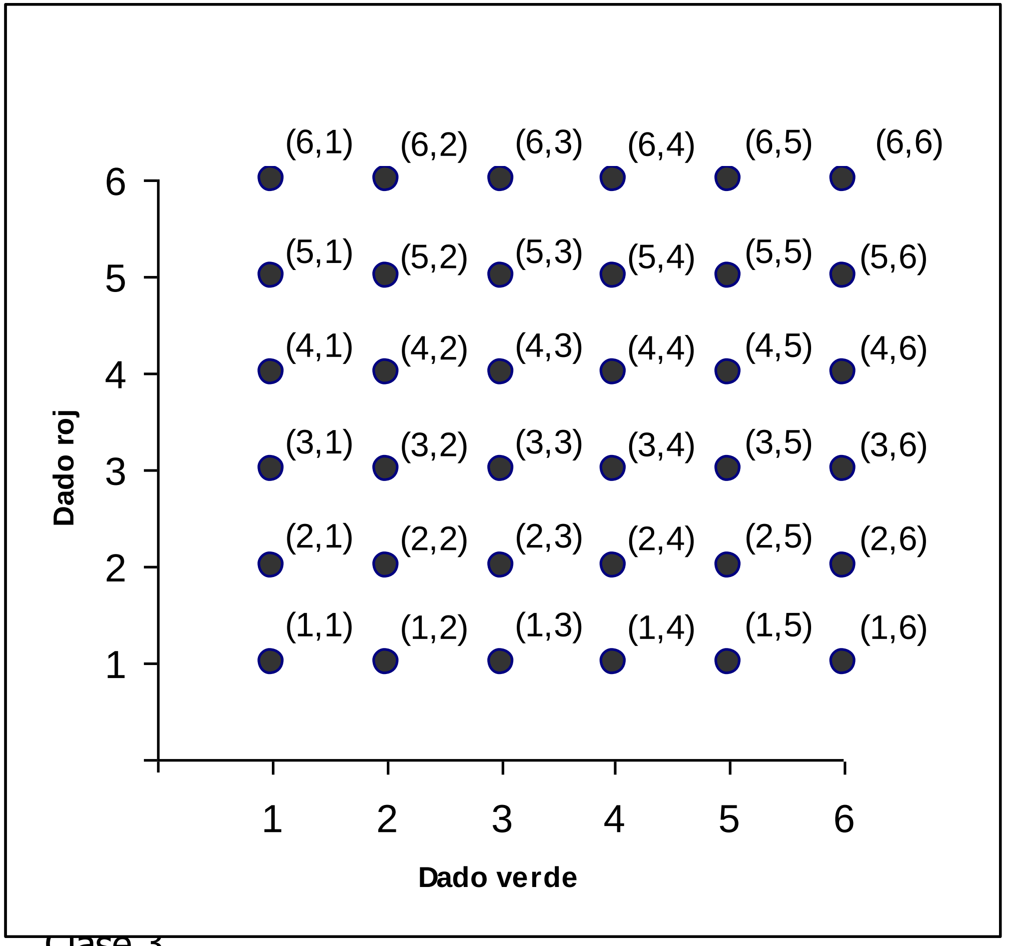
- **Mínimos cuadrados**
- **Máxima verosimilitud**
- **Métodos bayesianos**
- Método de momentos
- Minimizar el valor absoluto (regresión de cuantiles)

Motivación

– Al modelar los sistemas naturales debemos considerar que estos (modelos y/o datos) no son determinísticos.
– Si no más bien estocásticos.
– Esto significa que el sistema
  • tiene una variabilidad innata que no somos capaces de describir perfectamente
  • que no podemos observar perfectamente el sistema
  • Una combinación de ambos.
– modelo quedará totalmente especificado únicamente cuando definamos la parte estocástica de él.
– Esto involucra el uso de variables aleatorias.

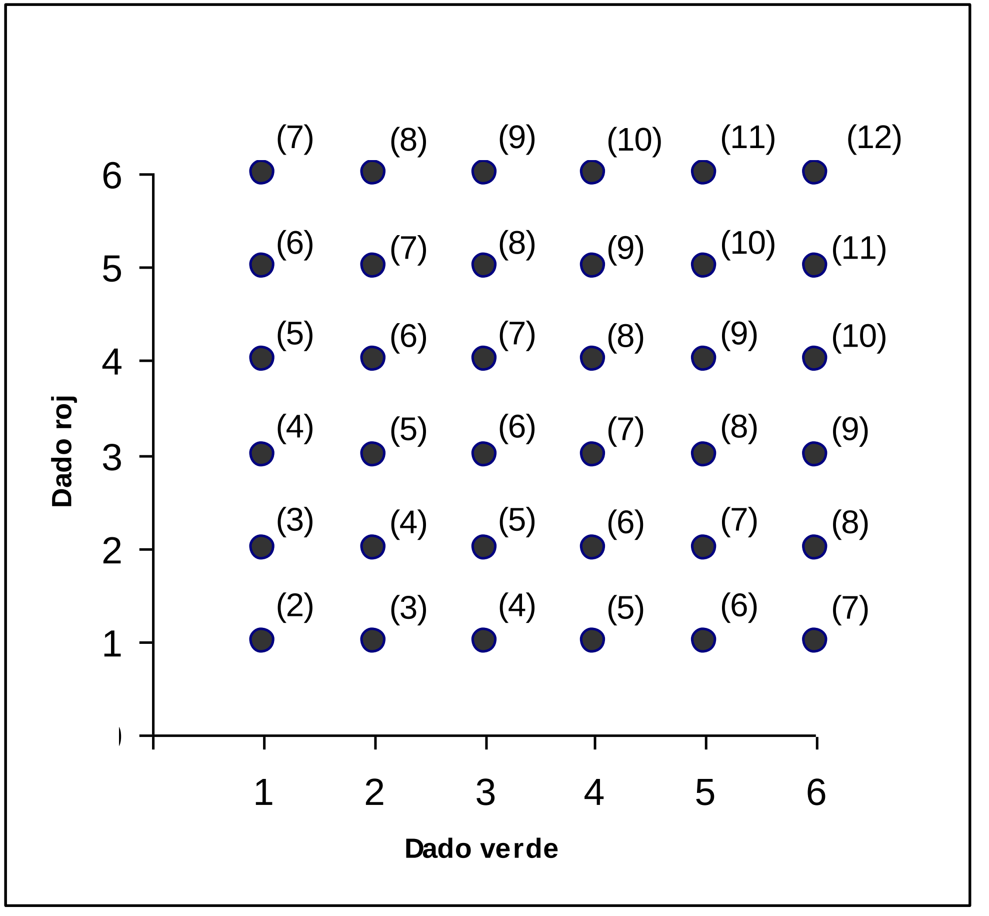
Probabilidades

- Supongamos que tiramos un dado verde y uno rojo una vez cada uno, el espacio muestral de este experimento es el siguiente



Probabilidades

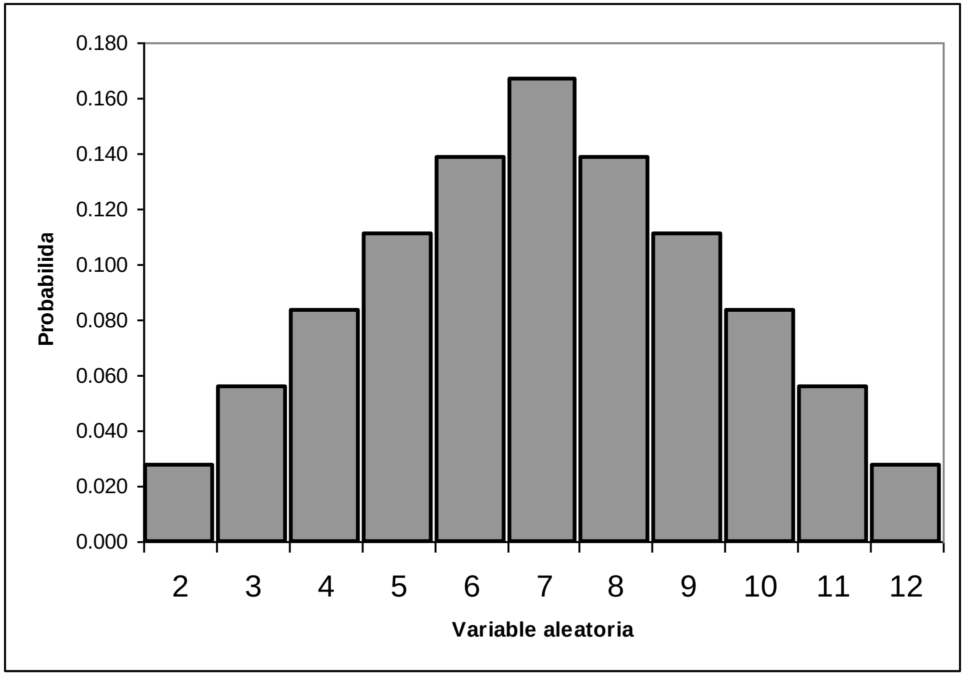
- Si los dados no estaban arreglados, entonces cada uno de los 36 resultados tiene probabilidad 1/36.
- Ahora si estamos interesados en conocer la suma de los dos valores de los dados (nuestra variable aleatoria), entonces los códigos (1,1), (1,2), (1,3)…(6,6) son reemplazados por:



    - Nuestra variable aleatoria puede adoptar valores entre 2 y 12, con probabilidad variable (no 1/36 en cada caso)

Distribución de probabilidades

- Al hablar de variables aleatorias → pensamos en distribuciones de probabilidades → asignan probabilidades a la variable aleatoria.
- En el ejemplo anterior la distribución de probabilidades es la siguiente



Distribución de probabilidades

- En vez de mostrar un gráfico con las probabilidades asociadas a la variable aleatoria es preferible usar una función f(x) o P(**x**=x) que permita calcular dichas probabilidades. En el ejemplo anterior podemos escribir:

Distribución de probabilidades

$$f(x) = \frac{6 - |x - 7|}{36}$$

$$f(2) = \frac{6 - |2 - 7|}{36} = \frac{1}{36}$$

$$f(7) = \frac{6 - |7 - 7|}{36} = \frac{6}{36}$$

- Una distribución de probabilidades debe satisfacer las siguientes condiciones:
- i) $f(x) \geq 0$, para cada valor dentro del dominio
- ii) $\sum_x f(x) = 1$, donde la sumatoria se extiende sobre todos los valores del dominio.

Distribuciones discretas de probabilidad

- Están asociadas a variables aleatorias discretas que pueden adoptar un número contable o finito de distintos valores. Algunas distribuciones de probabilidades discretas son:
  - Bernoulli
  - *Binomial*
  - Poisson
  - Geométrica

Binomial

- Ahora si repetimos el experimento múltiples veces, es decir colectamos muchas hembras en la población y determinamos su condición de madurez, entonces el número de hembras maduras será una variable aleatoria discreta con distribución Binomial:
- esta distribución tiene parámetros n y p, donde **n** es el número de veces que se realizó el experimento y **p** la probabilidad de que esté madura cada vez.
- La distribución de probabilidades es:

Binomial

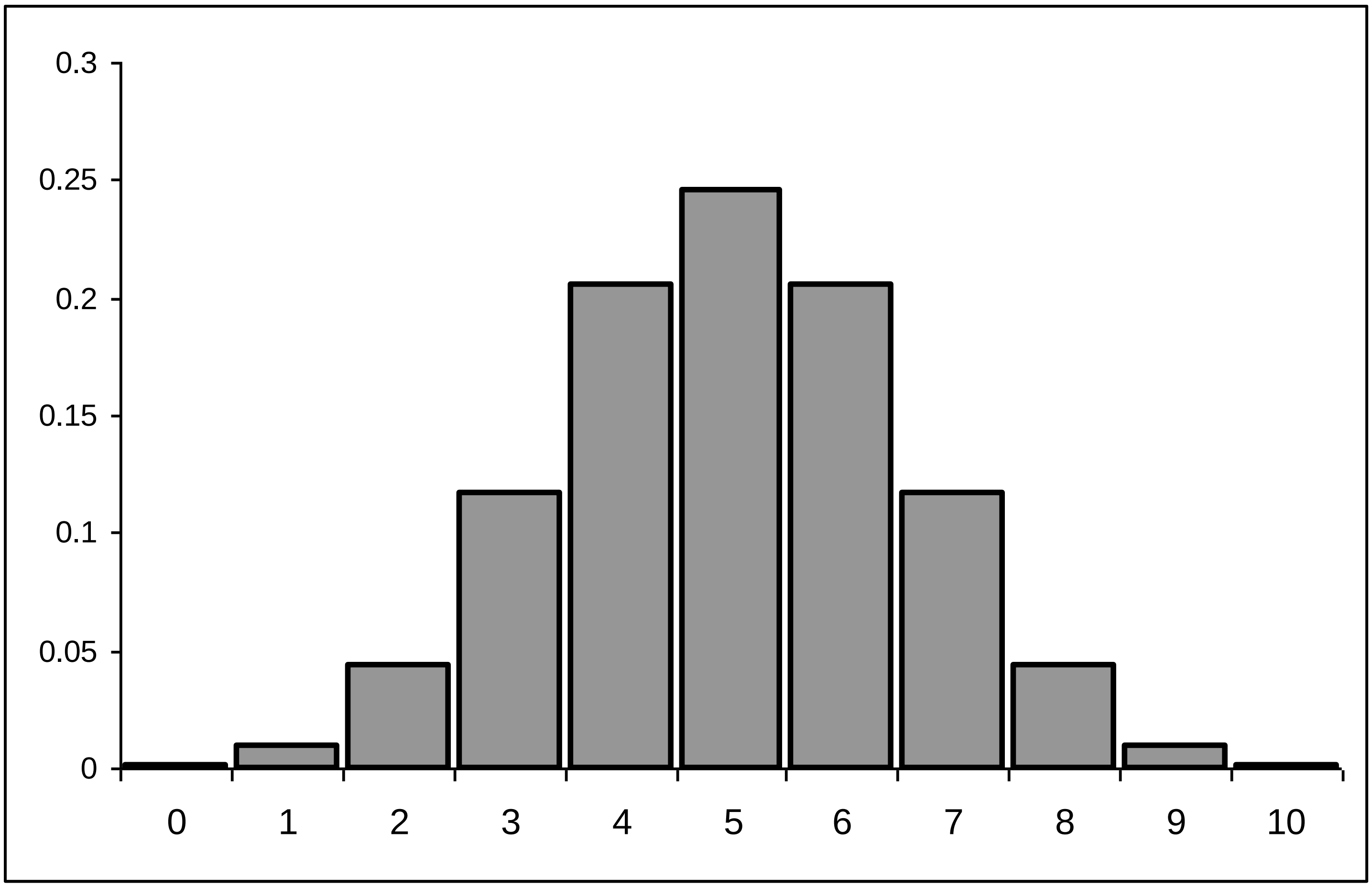
$$f(x) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Si p=0.5 y n=10, la probabilidad de que x=5 (5 cinco hembras estén maduras) es

$$f(x=5) = \binom{10}{5} 0.5^5 (1-0.5)^{10-5}$$

$$f(x=5) = 0.24$$

Binomial



Distribuciones continuas de probabilidad

- Cuando las variables aleatorias pueden adoptar un rango infinito de posibles valores (escala real) entonces estamos frente a variables continuas y sus distribuciones de probabilidad serán también continuas.
- Algunas distribuciones continuas de uso frecuente son:

Distribuciones continuas de probabilidad

- Uniforme
- Normal
- Lognormal
- Gamma
- Beta
- Chi-square

Distribución normal

- Es una distribución continua de mucha utilidad en probabilidad y estadística. Aparece frecuentemente tanto en estimación de parámetros como test de hipótesis. La función de densidad de probabilidades (pdf) es:

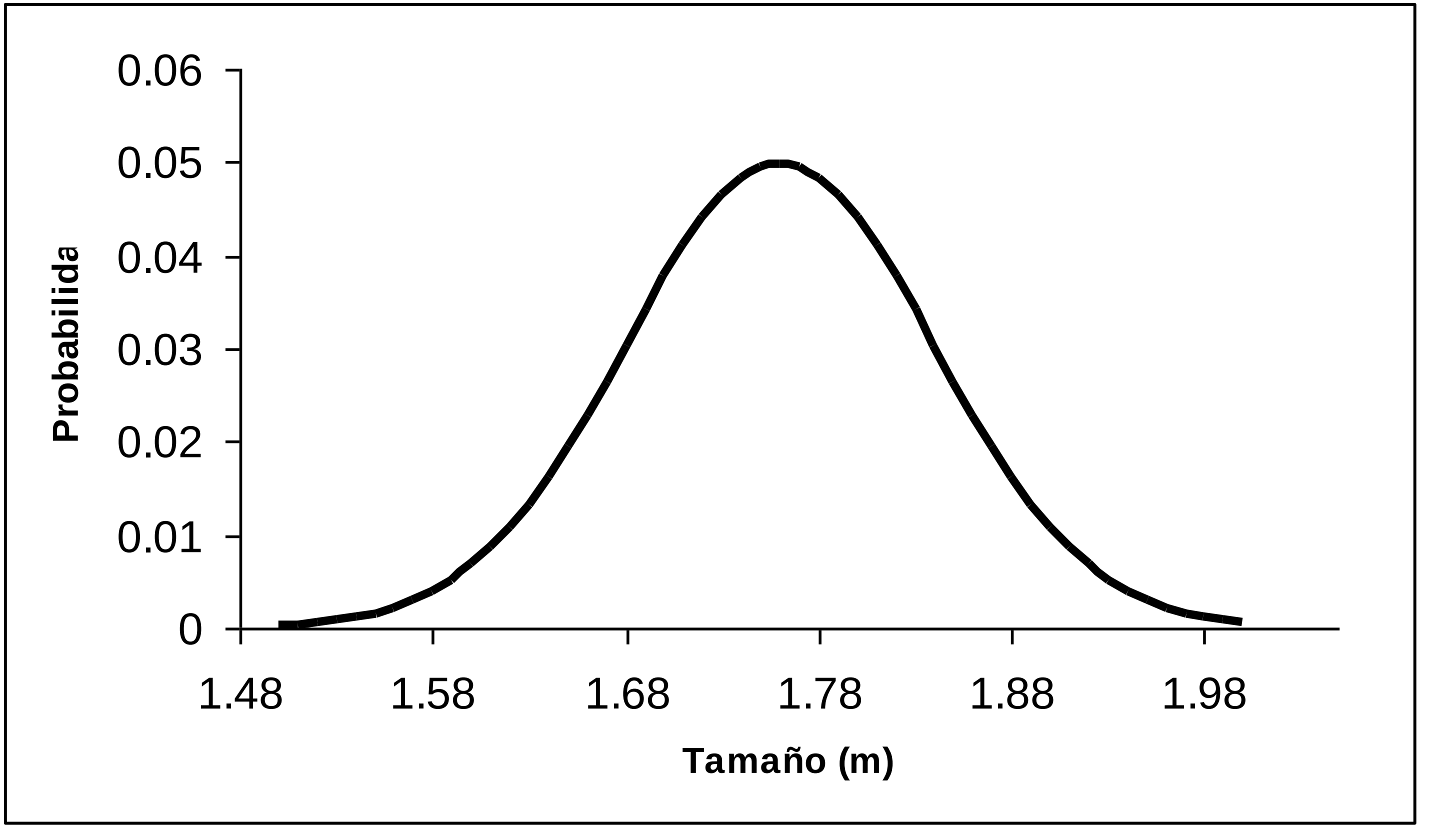
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ

Distribución normal

- Tiene dos parámetros: media (μ) y varianza (σ^2).
  $$x \sim N(\mu, \sigma^2)$$
- Si asumimos que nuestra variable aleatoria, digamos tamaño, tiene distribución normal con media 1.82 m y varianza (0.08^2), la representación gráfica de la distribución de probabilidades es

Distribución normal



Distribución normal

- NOTA 1: Si a nuestra variable aleatoria le restamos la media poblacional y dividimos esta diferencia por la desviación estándar (raíz cuadrada de la varianza) obtenemos una variable aleatoria normal estándar, con

$$\mu = 0 \quad , \quad \sigma^2 = 1$$

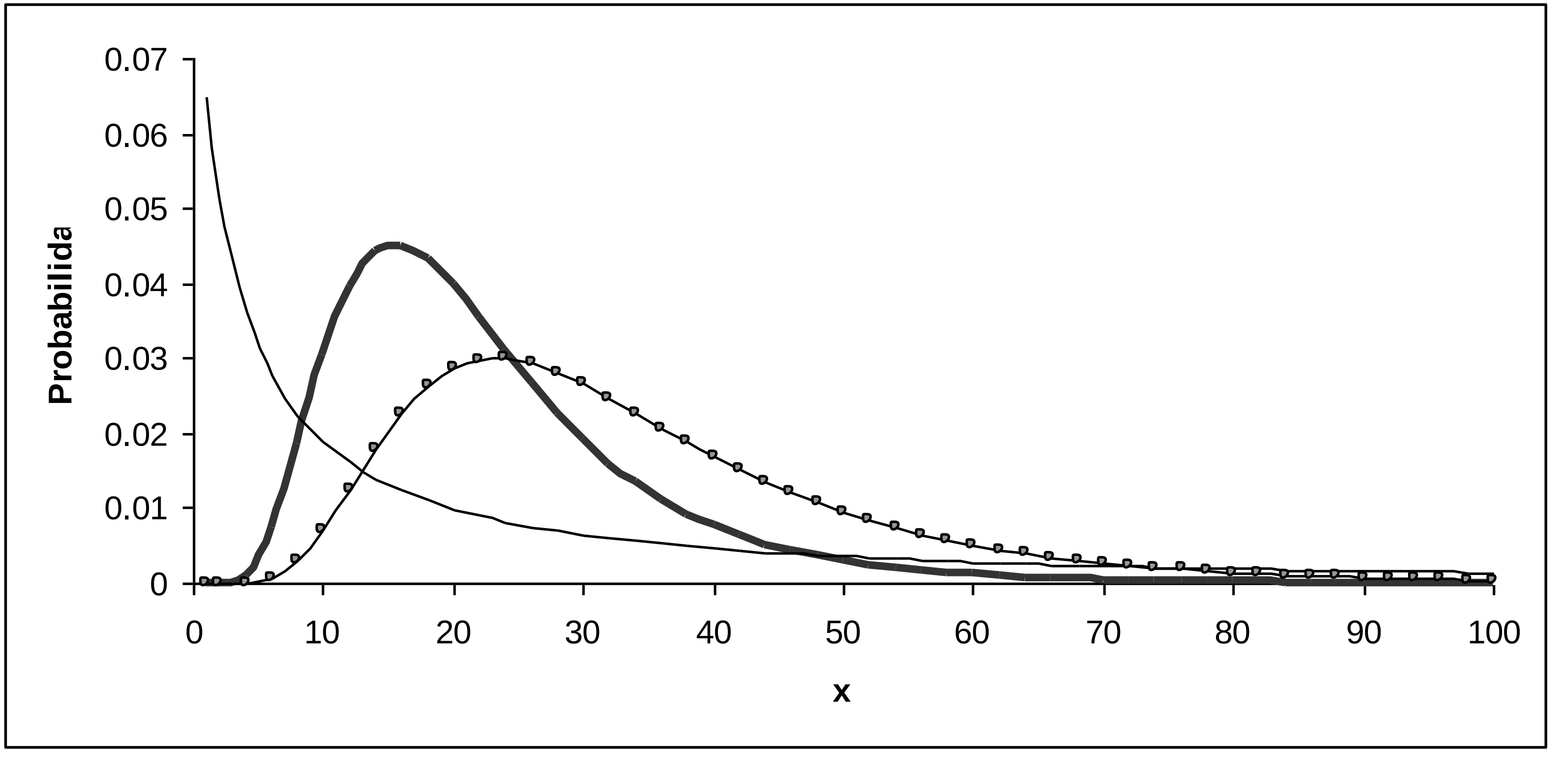
$$x \sim N(0,1)$$

Distribución Lognormal

- Es una distribución muy utilizada en estimación de parámetros de modelos de dinámica poblacional. Se utiliza a menudo para modelar la incertidumbre asociada a variables como abundancia y abundancia relativa. La pdf de una variable aleatoria lognormal es:

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(x)-\ln(\mu))^2}{2\sigma^2}}$$

Distribución Lognormal



Verosimilitud

- El renombrado matemático R.A. Fisher propuso en los años 1920s un método general de estimación de parámetros conocido como ***máxima verosimilitud***. Los estimadores obtenidos a través de este enfoque presentan muchas propiedades estadísticas deseables (mínima varianza, insesgados, etc).

Verosimilitud

- La característica esencial de este método es basar toda la inferencia en la información observacional o experimental recolectada (muestra aleatoria) y elegir como la mejor estimación del parámetro a aquel valor que maximice la probabilidad de haber observado la muestra. Si tomamos una muestra aleatoria la probabilidad de observarla se puede caracterizar a través de la siguiente probabilidad conjunta:

$$P(X_1 = x_{1,}\ X_2 = x_2, \ldots X_n = x_n) = f(x_1, x_2, \quad \ldots x_n; \theta)$$

Verosimilitud

- Una vez que se ha tomado la muestra, los x_i son constantes conocidas y la función $f(x_1, x_2 \ldots x_n; \theta)$ depende únicamente del vector de parámetros (θ). Esta función se le conoce como función de verosimilitud "L(θ)".
- Para obtener el valor del parámetro(s) que maximiza la función de verosimilitud se pueden usar métodos analíticos o numéricos:

Verosimilitud

- Analíticos: igualar la primera derivada parcial de L(θ) a cero y resolver el sistema de ecuaciones.
- Numéricos: usar un algoritmo de busqueda que maximice L(θ), variando θ .
- Si tratamos con modelos lineales en los parámetros preferiremos utilizar métodos analíticos para obtener estimadores de los parámetros. Si los modelos son de naturaleza no lineal solo podremos optar (a menos que sean linealizables) al uso de métodos numéricos.

Binomial

- Supongamos que tomamos una muestra de 89 hembras y 37 estaban maduras, cual es la proporción de madurez en la población de hembras?
- Esta variable se distribuye binomial → B(n,p). En este problema “n” es conocido (una observación) y “p” en cambio es un parámetro por estimar, dada la información disponible.
- Para estimar el parámetro (p) mediante el método de máxima verosimilitud debemos especificar la naturaleza de la variable aleatoria, que en este caso es una pmf (probability mass function) binomial:

Binomial

$$L(\theta) = b(x; n, \theta) = \binom{n}{x} \theta^{x} (1-\theta)^{n-x}$$

- Cual es el valor mas probable para θ si consideramos que la muestra aleatoria de 89 jaibas hembras 37 estaban maduras?

n = 89

x = 37

Binomial

- Si reemplazamos **n** y **x** en la ecuación anterior y le asignamos distintos valores a θ, entonces podremos evaluar cual es la verosimilitud asociada a cada valor del parámetro. Si repetimos esto en el rango del parámetro (entre 0 y 1, ya que es una probabilidad) estaremos construyendo un perfil de verosimilitud de θ. Según este método estadístico nos quedaremos con aquel valor del parámetro donde la verosimilitud sea máxima

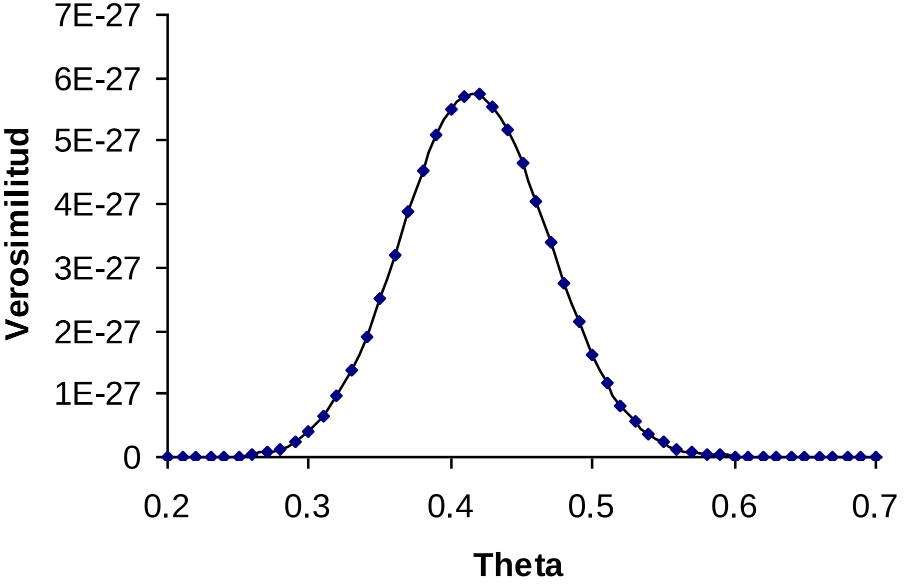
Binomial

- Haciendo una inspección visual sobre la tabla de valores calculados identificamos al valor del parámetro igual a 0.42 como al más verosímil.

Theta	**Verosimilitud**
0.3	3.97E-28
0.31	6.32E-28
0.32	9.57E-28
0.33	1.38E-27
0.34	1.91E-27
0.35	2.52E-27
0.36	3.20E-27
0.37	3.88E-27
0.38	4.53E-27
0.39	5.09E-27
0.4	5.50E-27
0.41	5.72E-27
0.42	**5.73E-27**
0.43	5.54E-27
0.44	5.17E-27
0.45	4.65E-27
0.46	4.04E-27
0.47	3.39E-27
0.48	2.74E-27
0.49	2.14E-27
0.5	1.62E-27
0.51	1.18E-27

Binomial

- En este caso podemos obtener un estimador analítico si optimizamos mediante cálculo la función objetivo:



Binomial

$$L(\theta) = \binom{n}{x} \theta^{x} (1-\theta)^{n-x}$$

$$\log L(\theta) = \log\left(\binom{n}{x} \theta^{x} (1-\theta)^{n-x} \right)$$

$$\log L(\theta) = \log\binom{n}{x} + x\log(\theta) + (n-x)\log(1-\theta)$$

$$\partial \log L / \partial \theta = \log\binom{n}{x} + x\,\partial \log(\theta)/\partial \theta + (n-x)\,\partial \log(1-\theta)/\partial \theta = 0$$

Binomial

$$\partial \log L / \partial \theta = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$$

$$x(1-\theta) = (n-x)\theta$$

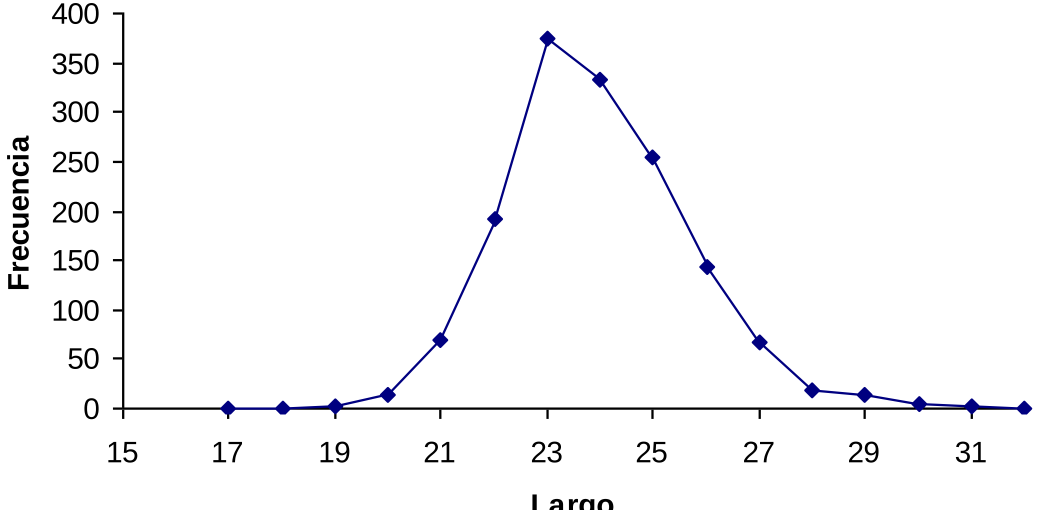
$$\theta = \frac{x}{n}$$

$$\theta = \frac{37}{89} = 0.4157$$

- Donde el cuociente es el estimador de θ .
- Reemplazando obtenemos la estimación puntual del parámetro:

Normal

- Cálculo del promedio. Supongamos que recolectamos datos de longitud total de la presa de un mamífero marino. La distribución de frecuencia se observa en el siguiente gráfico.



Normal

- Nos interesa estimar la talla promedio utilizando el método de máxima verosimilitud, por lo tanto debemos especificar la distribución de probabilidades asociada a nuestra variable aleatoria, que en este caso será normal. La verosimilitud de cada observación esta dada por la siguiente expresión:

$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

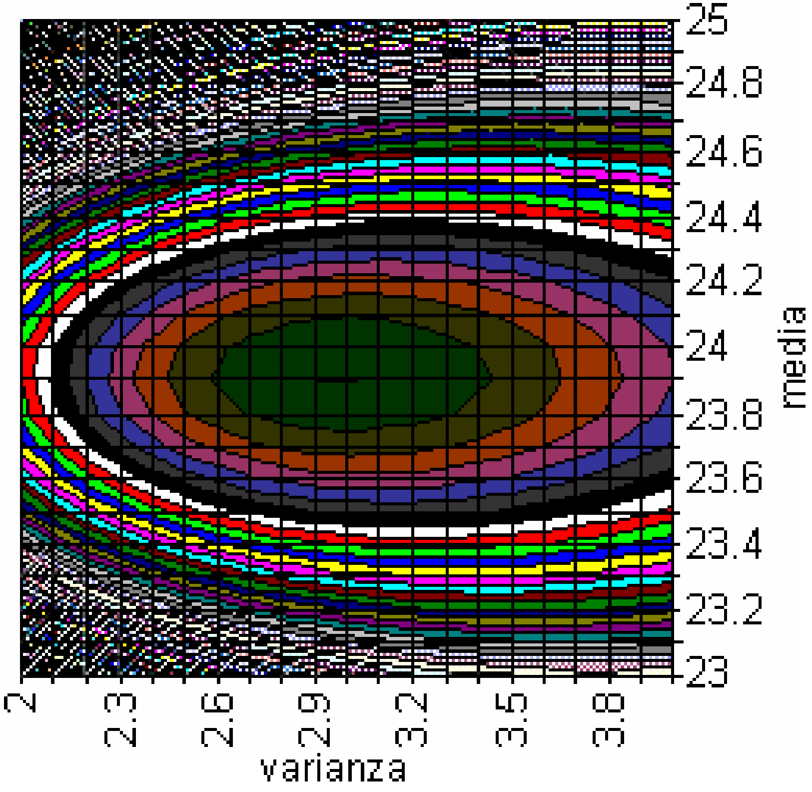
Normal

- Donde la verosimilitud es función de dos parámetros, μ (media) y σ^2 (varianza).
- Ahora la verosimilitud de todas las observaciones (independientes) estará determinada por

$$L = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

Normal

- A diferencia del problema (a), en este caso la función de verosimilitud depende de 2 parámetros, de tal forma que para graficar la verosimilitud debemos hacer un gráfico bidimensional (media y varianza en cada eje).



Normal

- Este gráfico nos describe la forma la superficie de verosimilitud indicando:
  - Que la máxima verosimilitud se encuentra con una media de ca. 23.9 y una varianza de 3.
  - Prácticamente no existe correlación entre los parámetros.
- Un problema de esta naturaleza puede ser abordado en forma analítica de tal forma que encontremos estimadores de máxima verosimilitud para μ y σ^2.

Normal

$$L = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\log L = -\frac{n}{2}\log\left[2\pi\right] - \frac{n}{2}\log\left[\sigma^2\right] - \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$$

$$-\log L = \frac{n}{2}\log\left[\sigma^2\right] + \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$$

- Aplicamos la derivada parcial de –log(L) con respecto a , igualamos a cero y resolvemos para este parámetro

Normal

$$-\frac{\partial \log L}{\partial \mu} = \frac{\left[\sum_i (x_i - \mu)^2\right]'}{2\sigma^2} = 0$$

$$\left[\sum_i (x_i - \mu)^2\right]' = 0$$

$$\left[\sum_i x_i^{\,2} - 2\mu \sum_i x_i + \sum_i \mu^2\right]' = 0$$

$$-2\sum_i x_i + 2n\mu = 0$$

$$\mu = \frac{\sum_i x_i}{n}$$

- Siendo el cuociente el estimador de la media.

Normal

$$-\log L = \frac{n}{2}\log\left[\sigma^2\right] + \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$$

- Para obtener el estimador analítico de σ^2 aplicamos la derivada parcial de –log(L) con respecto a σ^2, igualamos a cero y resolvemos para este parámetro

Verosimilitud conjunta: Combinando varias fuentes de información

- Se comparten parámetros entre distintos sets de datos
- Se estiman los parámetros maximizando la verosimilitud total (asumiendo independencia)

$$L(\boldsymbol{\theta} \mid \mathbf{data}) = \prod_i L(\boldsymbol{\theta} \mid \mathrm{data}_i)$$

Se pueden combinar diversas funciones de versomilitud

Verosimilitud conjunta

- Se reparametrizan los modelos para que compartan parámetros de los mismos procesos
- Se estiman todos los parámetros en forma simultánea al maximizar la verosimilitud conjunta
- Los parámetros compartidos tienen el mismo valor a través del análisis

$$L(\boldsymbol{\theta} \mid \mathbf{data}) = L(\boldsymbol{\theta} \mid \mathrm{data}_A) L(\boldsymbol{\theta} \mid \mathrm{data}_B)$$

$$L(\boldsymbol{\theta} \mid \mathrm{data}_A) = f_A(\mathrm{data}_\mathrm{A}, \boldsymbol{\varphi}, \boldsymbol{\eta}) \qquad L(\boldsymbol{\theta} \mid \mathrm{data}_B) = f_B(\mathrm{data}_\mathrm{B}, \boldsymbol{\kappa}, \boldsymbol{\eta})$$

$$\boldsymbol{\theta} = \{\boldsymbol{\varphi}, \boldsymbol{\kappa}, \boldsymbol{\eta}\}$$

Parámetros compartidos $\boldsymbol{\eta}$

Algunas dificultades

- Más datos no es necesariamente mejor, si el modelo esta mal especificado
  - E.g. Los datos de composición de edades dificultarán la estimación de tendencias de abundancia, si las selectividades están mal especificadas
- Especificación arbitraria de los pesos asignados a cada fuente de información

Modelo de Observación

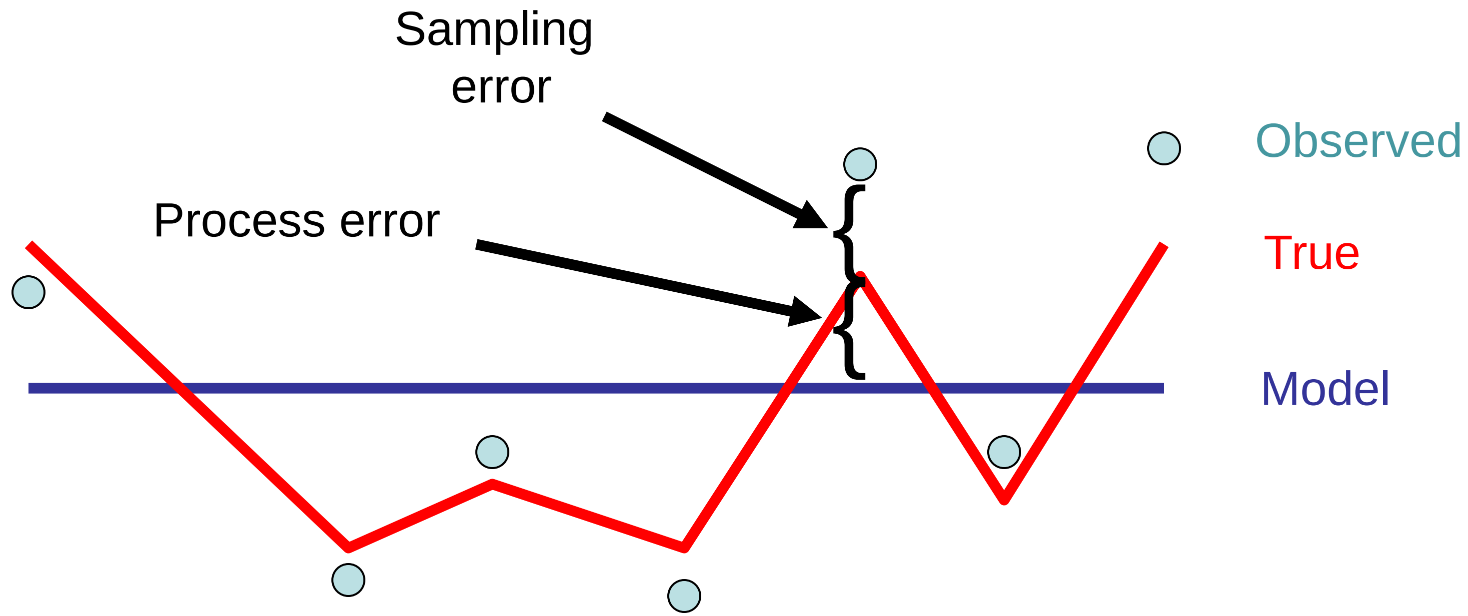
- El modelo de observación puede no ser función de una medición directa de una predicción principal del modelo (abundancia)

    E.g. Indice de abundancia

    - $I = q*B$

- La verosimilitud puede no utilizar una predicción directa de la observación
  - E.g. composiciones de edad
    - MN(Proporciones, tamaño de muestra)

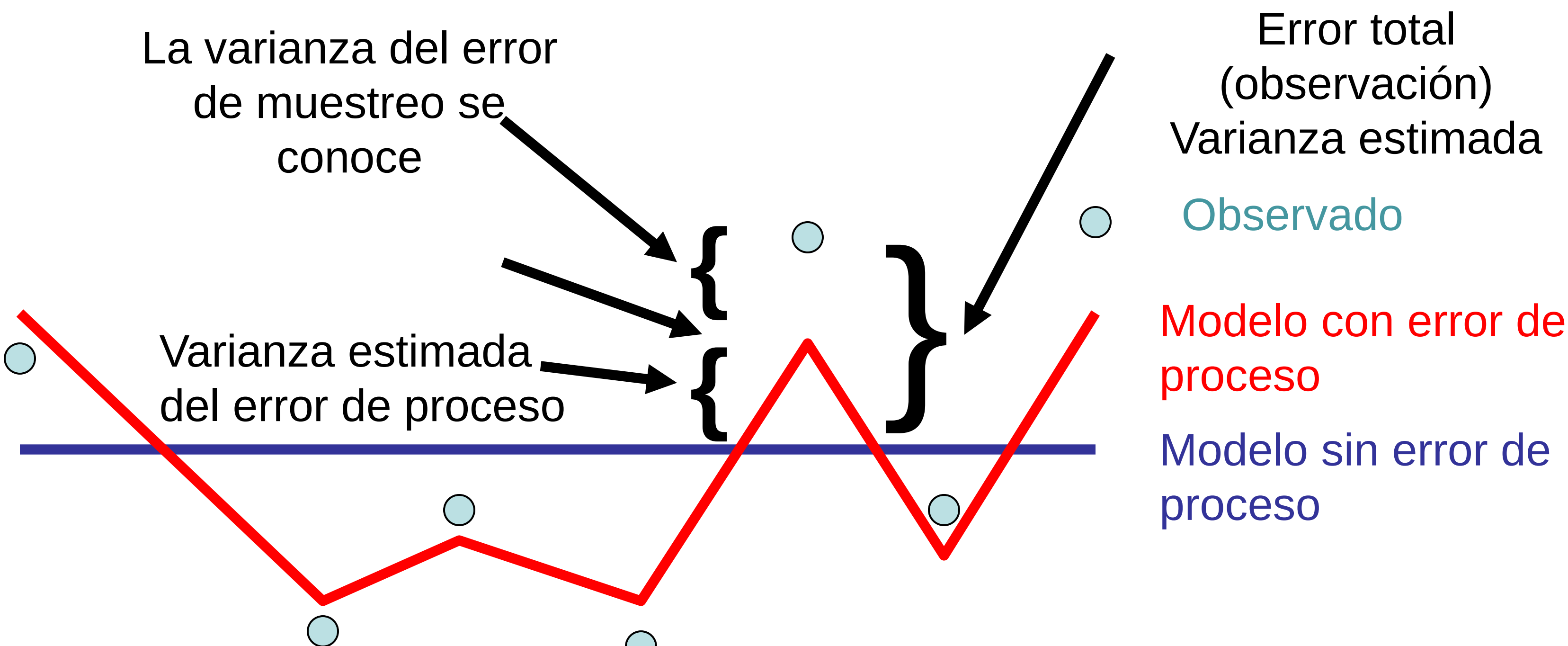
Error Proceso / observación



Conozca su error de muestreo

- A menudo podemos estimar nuestro error de muestreo
- Conocer el error de muestreo nos permite separar error de proceso y de observación
- Modelar apropiadamente los errores es importante

Error de proceso/variabilidad



Complicaciones en la calibración de modelos

- Alta correlación entre parámetros
- Problemas con derivadas numéricas
- Problema de discontinuidad de las funciones
- Parámetros enteros
- Múltiples mínimos
- Espacio paramétrico restringido

Non-linear minimization

- La optimización no lineal tiene tanto de arte como de ciencia
- No se puede tan solo meter algunos valores en el programa y esperar que lo mejor resulte
  - Se debe comenzar de valores iniciales razonables
  - Plotee los datos utilizando los valores iniciales
  - Verifique convergencia
- Proceso toma tiempo y experiencia.

Verosimilitud normal

Verosimilitud (Likelihood)

$$L(\theta \mid data) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(O-P\right)^2}{2\sigma^2} \right]$$

-ln(Likelihood)

$$-\ln L(\theta \mid data) = 0.5\ln\left(2\pi\right) + \ln\left(\sigma\right) + \frac{\left(O-P\right)^2}{2\sigma^2}$$

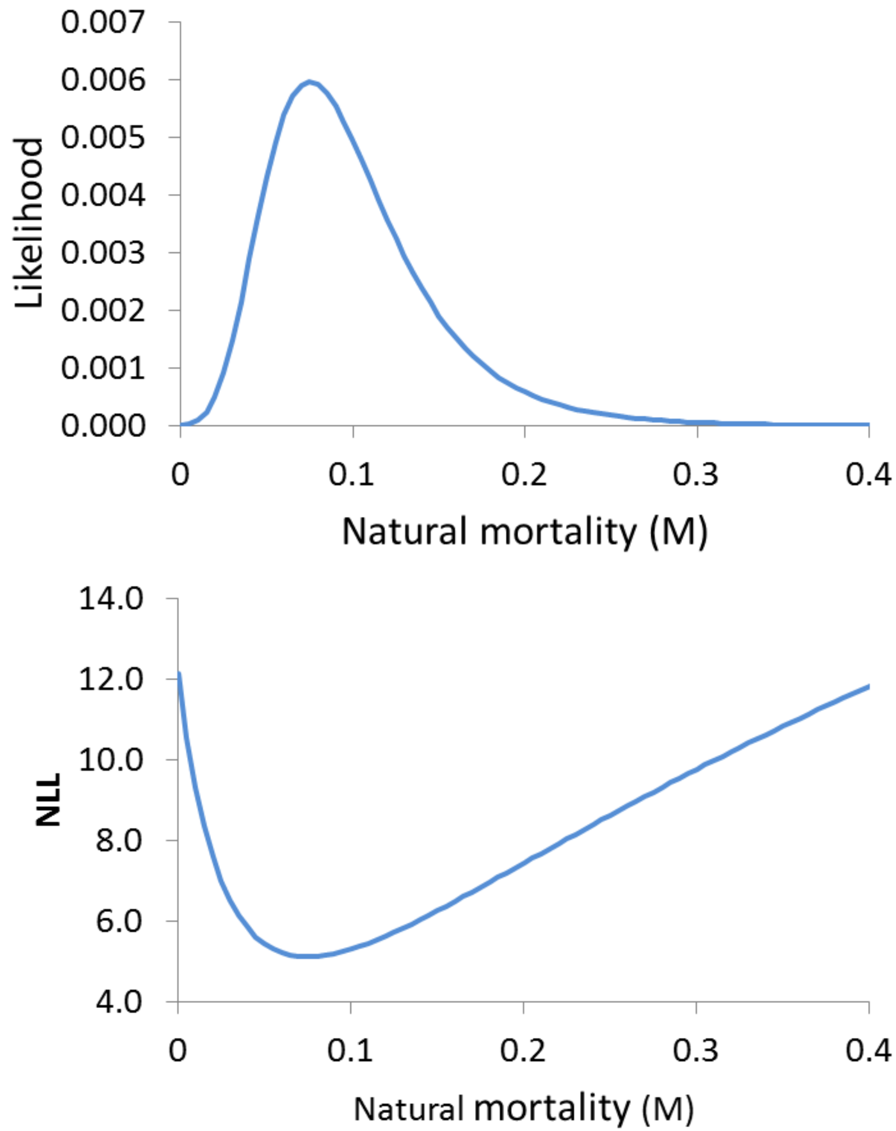
-ln(Likelihood) sin constantes

$$-\ln L(\theta \mid data) = \ln\left(\sigma\right) + \frac{\left(O-P\right)^2}{2\sigma^2}$$

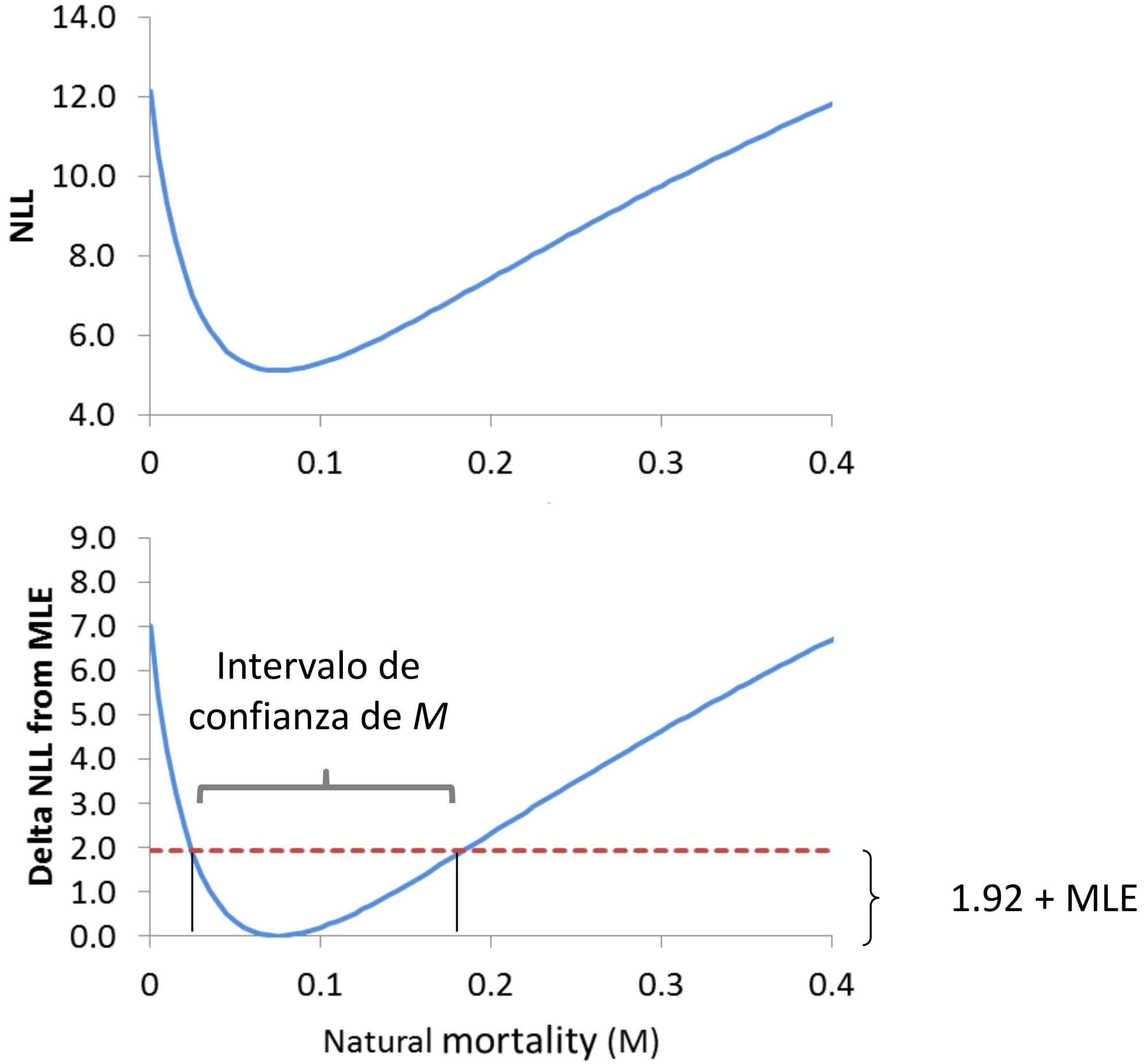
-ln(Likelihood) sin constantes, σ conocido

$$-\ln L(\theta \mid data) = \frac{\left(O-P\right)^2}{2\sigma^2}$$

Likelihood y -log(Likelihood)



Perfil de verosimilitud e intervalos de confianza



Análisis integrado contemporaneo de evaluación de stock

- No esta restringido a lo que se habia hecho en el pasado
  - e.g. No se requiere algebra para estimar los parámetros
- Habilidad para utilizar la informacion disponible
  - e.g. reformule el modelo para utilizar la data disponible
  - use distribuciones a priori
- Sea consistente, no ad-hoc
  - e.g. Análisis integrado asegura que los supuestos se utilizan en forma consistente a través de todo el análisis
  - Use aproximaciones estadísticas y represente la incertidumbre
  - e.g. Verosimilitud y métodos bayesianos
- Extraiga toda la información disponible de los datos
  - e.g. Al integrar la información de marcaje y recaptura en la evaluación de stock para obtener información de abundancia, mortalidad natural, selectividad, movimiento

Importante

- La varianza y el tamaño de muestra controlan la influencia de la data