Agua/Tierra

Tomado de: Statistical Rethinking –
Richard McElreath, Capítulos 2 y 3

Agua/Tierra

- Hay un planeta con una atmósfera tan densa que no nos permite ver su superficie
- Queremos saber la cobertura de agua en ese planeta
- Enviamos sondas que aterrizan de manera aleatoria sobre la superficie del planeta
- Cada sonda nos informa si aterriza sobre agua o sobre tierra

- ¿Podemos inferir la proporción de cobertura de agua de el planeta con sólo pocas medidas?

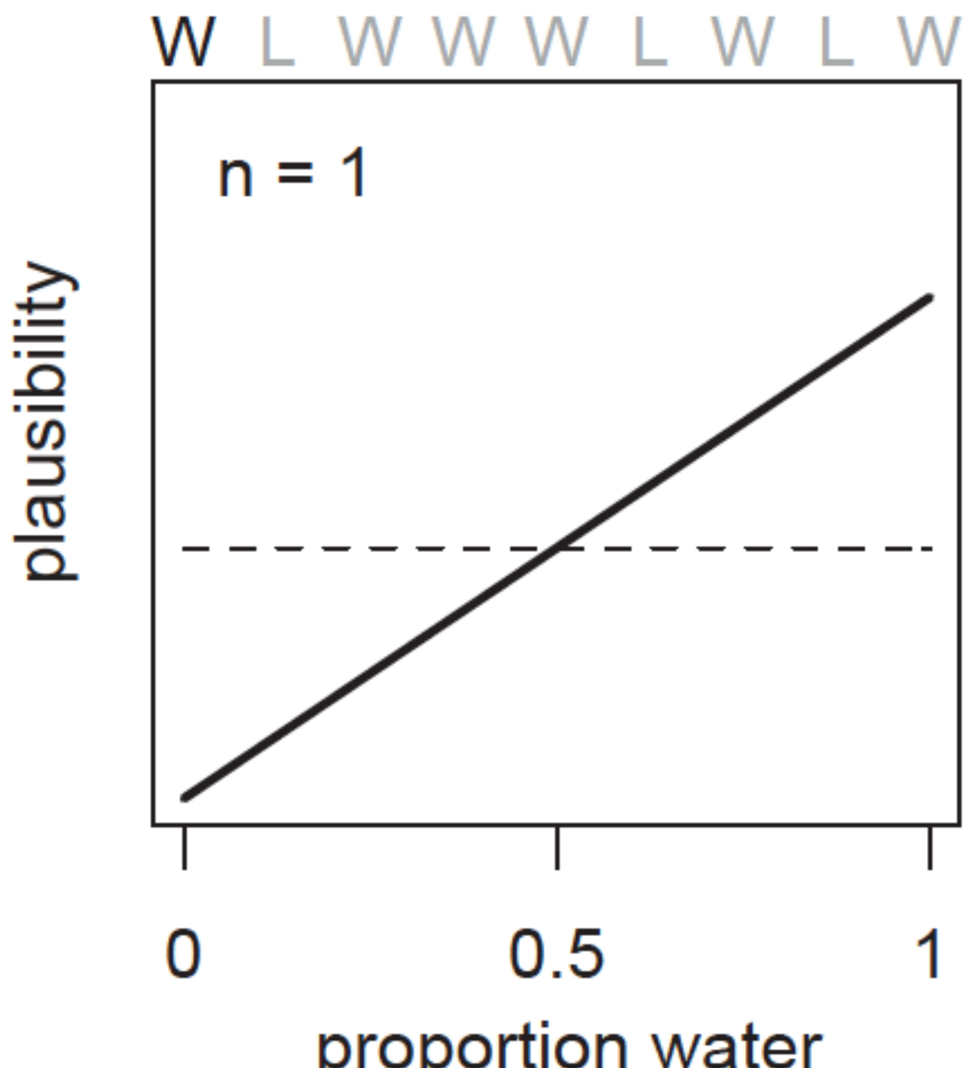
Agua/Tierra

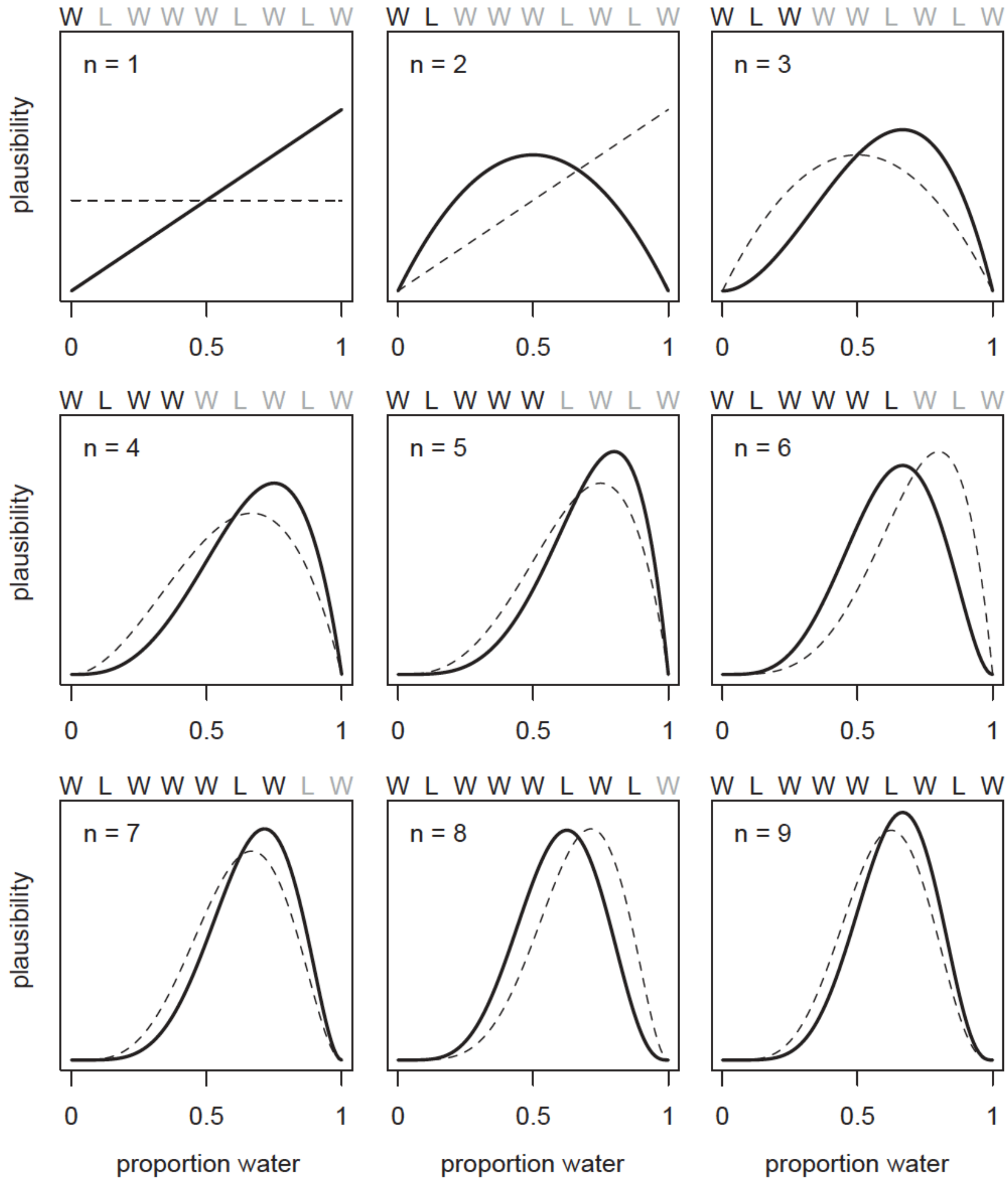
Datos: W L W W W L W L W

Historia para los datos

- (1) La verdadera proporción de superficie cubierta por agua es p
- (2) Cada sonda que aterriza tiene una probabilidad p de producir una observación de agua (W) y una probabilidad $1\text{-}p$ de producir una observación de tierra (L)
- (3) Cada aterrizaje de sonda es independiente de los otros

Actualizar





W L W W W L W L W
n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
n = 7
n = 8
n = 9
plausibility
0
0.5
1
proportion water

El proceso de inferencia

- (1) El número de maneras en que cada conjetura puede producir una observación
- (2) El número acumulado de maneras en que cada conjetura puede producir todos los datos
- (3) La plausibilidad inicial para cada conjetura

Verosimilitud

- Escogemos una expresión matemática que pueda explicar (generar) las observaciones
- En este caso hay dos opciones para cada dato
- Cada sonda es independiente de las otras
- La probabilidad p de observar W es la misma en todos los lanzamientos

-> Distribución binomial

Verosimilitud

- Probabilidad de que dado un valor de p, haya un número w de observaciones de W en n lanzamientos

$$\Pr(w|n,p) = \frac{n!}{w!(n-w)!}p^w(1-p)^{n-w}$$

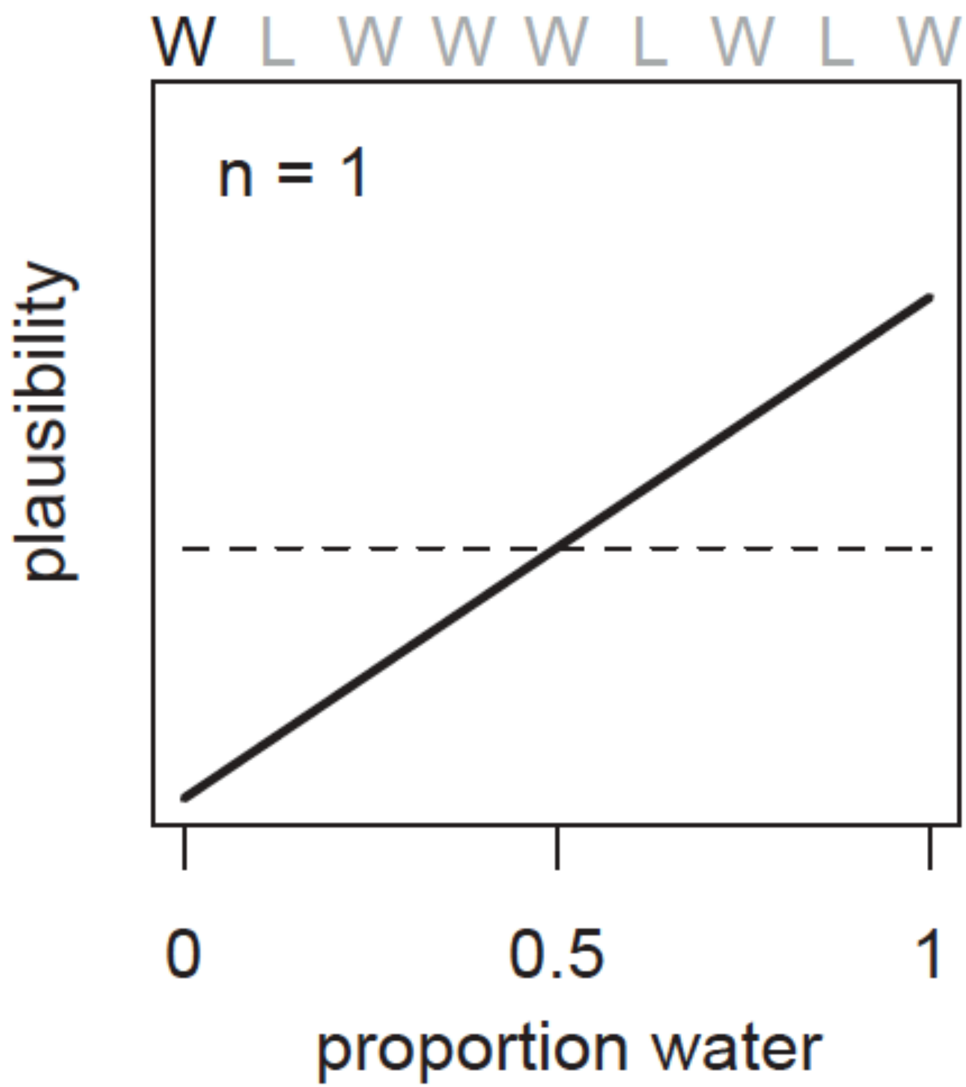
Prior

- Todos los valores en el rango [0,1] son igualmente probables

$$\Pr(p) = \frac{1}{1-0} = 1$$

- *Prior débilmente informativo/a*

Prior: Línea punteada



Posterior

- Objetivo: Dados los datos (!) ¿cuál es la probabilidad de que el parámetro tenga cierto valor?

$$\Pr(p|n, w)$$

- Regla de Bayes (obviando n):

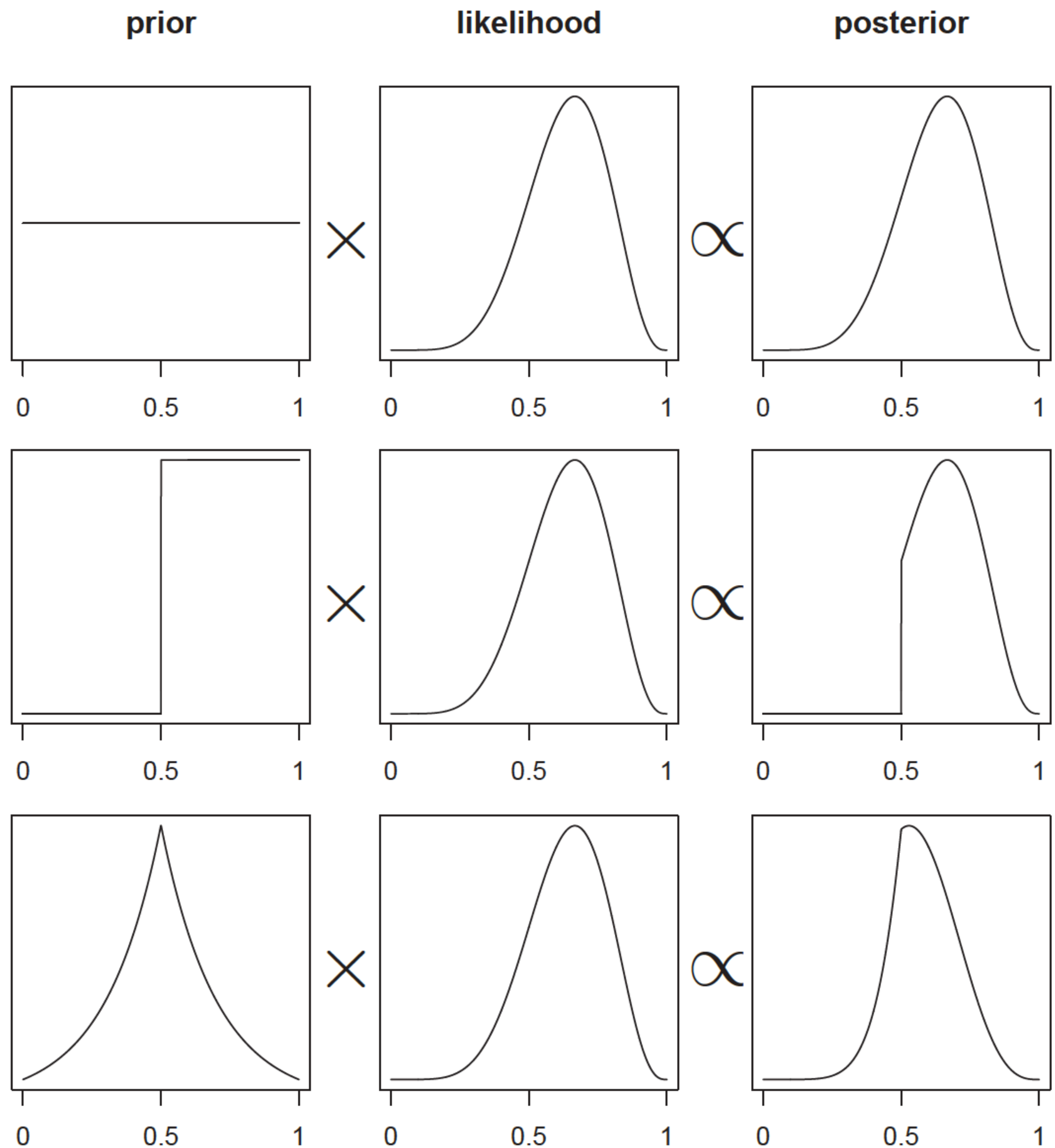
$$\Pr(p|w) = \frac{\Pr(w|p)\Pr(p)}{\Pr(w)}$$

Posterior

$$\text{Posterior} = \frac{\text{Verosimilitud} \times \text{Prior}}{\text{Verosimilitud promedio}}$$

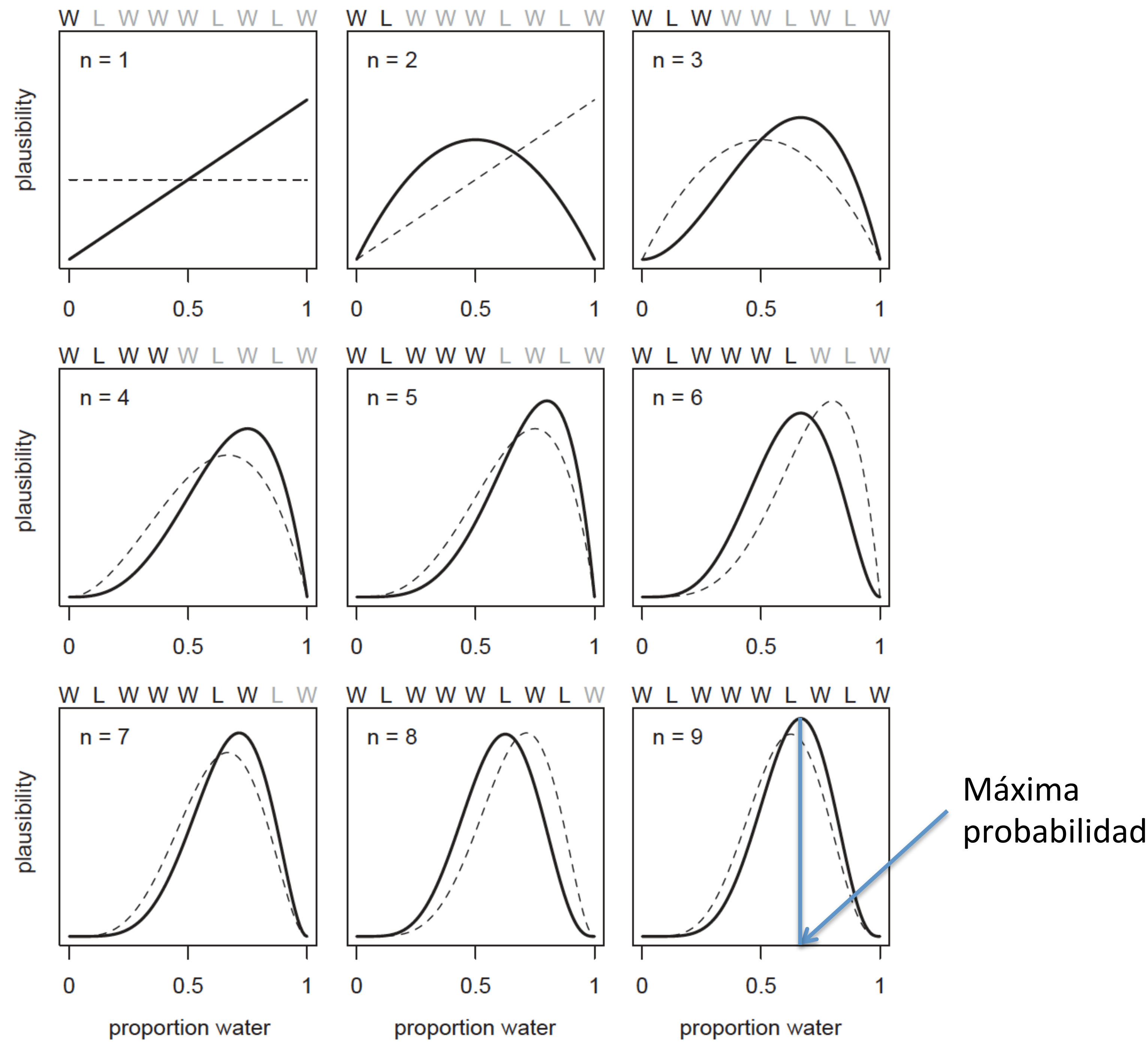
$$\Pr(w) = \mathrm{E}\big(\Pr(w|p)\big) = \int \Pr(w|p)\Pr(p)dp$$

Sirve para que la probabilidad posterior sume 1


prior
likelihood
posterior
0
0.5
1

Estimación de Posterior

- Grid (fuerza bruta)
- Maximum Likelihood Estimation (frecuentista)
- Monte Carlo (eficiente)
- Markov Chain Monte Carlo (muy eficiente)



W L W W W L W L W
n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
n = 7
n = 8
n = 9
plausibility
0
0.5
1
proportion water
Máxima
probabilidad

Máxima Verosimilitud

- Aproximación no bayesiana (frecuentista)
- Optimización (hallar el máx/min de la posterior)
- Valor óptimo p = 2/3
- ¿Por qué?

Andr(é|ew) MacDonald
@polesasunder

Follow

build an understanding of statistics and you too can cripple you publication record while also alienating all your colleagues

7:43 PM - 3 Apr 2019

193 Retweets **1,419** Likes

26 193 1.4K

Recommended steps in the statistical analysis of scientific data

- exploration of the data
- careful statement of the scientific problem
- model formulation in mathematical form
- choice of statistical method(s)
- calculation of statistical quantities
- judicious scientific evaluation of the results

Statistics: Some basic definitions

- Statistical inference
  - Seeking quantitative insight & interpretation of a dataset
- Hypothesis testing
  - To what confidence is a dataset consistent with a previously stated hypothesis?
- Estimation
  - Seeking the quantitative characteristics of a functional model designed to explain a dataset. An estimator seeks to approximate the unknown parameters based on the data
- Probability distribution
  - A parametric functional family describing the behavior of a parent distribution of a dataset (e.g. Gaussian = normal)
- Nonparametric statistics
  - Inference based directly on the dataset without parametric models Independent & identically distributed (iid) data point
  - A sample of similarly but independently acquired quantitative measurements.

- Frequentist statistics
  - Suite of classical inference methods based on simple probability distributions. Fixed hypotheses.
- Bayesian statistics
  - Inference methods based on Bayes' Theorem based on likelihoods and prior distributions. Changing hypotheses.
- L_1 and L_2 methods
  - 19th century methods for estimation based on minimizing the absolute or squared deviations between a sample and a model.
- Maximum likelihood methods
  - 20th century methods for parametric estimation based on the likelihood that a dataset fits the model (often like L_2).
- Gibbs sampling, Metropolis-Hastings algorithm, Markov chain Monte Carlo, …
  - New computational methods useful for integrations over hypothesis space in Bayesian statistics.

- Robust (nonparametric) methods
  - Statistical procedures that are insensitive to data outliers or distributions
- Model selection & validation
  - Procedures for estimating the goodness-of-fit and choice of parametric model. (Nested vs. non-nested models, model misspecification)
- Statistical power, efficiency & bias
  - Mathematical evaluation of the effectiveness of a statistical procedure to achieve its desired goals
- Two-sample & k-sample tests
  - Statistical tests giving probabilities that k samples are drawn from the same parent sample
- Independent & identically distributed (i.i.d.) data point
  - A sample of similarly but independently acquired quantitative measurements.
- Heteroscedasticity
  - A failure of i.i.d. due to differently weighted data points, common in astronomy due to measurement errors with known variances

Applied statistics methods

- Multivariate analysis
  - Establishing the structure of a table of rows & columns
  - Analysis of variance, regression, principal component analysis, discriminant analysis, factor analysis
- Multivariate classification
  - Dividing a multivariate dataset into distinct classes
- Correlation & regression
  - Establishing the relationships between variables in a sample
- Time series analysis
  - Studying data measured along a time-like axis
- Spatial analysis
  - Studying point or continuous processes in 2-3-dimensions
- Survival analysis
  - Studying data subject to censoring (e.g. upper limits)
- Data mining
  - Studying structures in mega-datasets