laSalle
UNIVERSITAT RAMON LLULL

Bayes and Friends

Máster Universitario en Ciencia de los Datos / Data Science (MUDS)

Ángel Berián, Ricard Sierra, Xavier Vilasís

Objetivos de Aprendizaje

Interpretación

Interpretar la probabilidad como un grado de creencia en lugar de frecuencia pura.

Actualización

Aplicar el Teorema de Bayes para actualizar creencias a la luz de nueva evidencia.

Inferencia

Construir inferencia desde la posterior: estimación puntual e intervalos creíbles.

Motivación: "Aprender con Evidencia"

La Idea Central:

- El aprendizaje no es estático; es un proceso continuo de actualización.
- Nuestras creencias (priors) cambian cuando observamos nuevos datos (evidencia).

Casos de Uso en Data Science:

- **A/B Testing:** Decisiones más rápidas con muestras pequeñas.
- **Riesgo:** Modelado de eventos raros.
- **Medicina:** Diagnóstico basado en prevalencia y síntomas.

Thomas Bayes

Dos Paradigmas en Pugna

Concepto	Frecuentista (Clásico)	Bayesiano
Probabilidad	Frecuencia a largo plazo en experimentos repetidos.	Grado de creencia o certeza subjetiva.
Parámetros	Fijos y desconocidos.	Variables aleatorias (incertidumbre).
Salidas	P-values, Intervalos de Confianza.	Distribución Posterior, Intervalos de Credibilidad.

Probabilidad Subjetiva

¿Por qué θ (theta) es aleatoria?

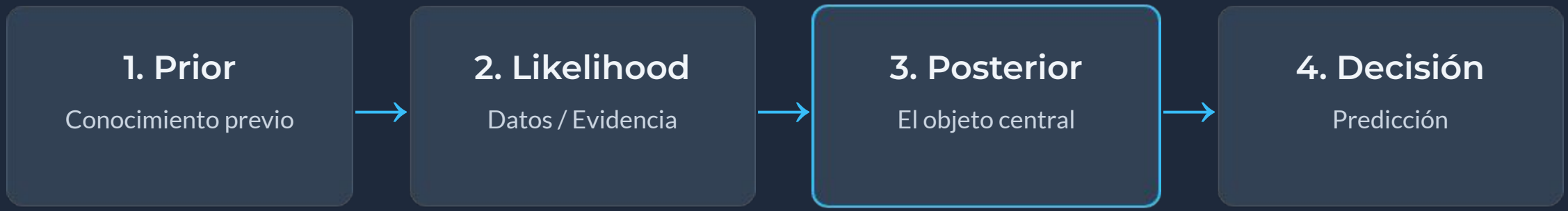
En el mundo Bayesiano, no decimos que el parámetro físico cambia, sino que nuestra **incertidumbre** sobre él se modela como una distribución.

> Inferencia = Distribución Completa

No buscamos solo un número (estimación puntual), buscamos la forma completa de la incertidumbre.

"En el contexto de la estadística bayesiana, la letra griega θ (theta) representa el parámetro o conjunto de parámetros desconocidos que intentamos estimar a partir de los datos observados"

Flujo de Trabajo Bayesiano



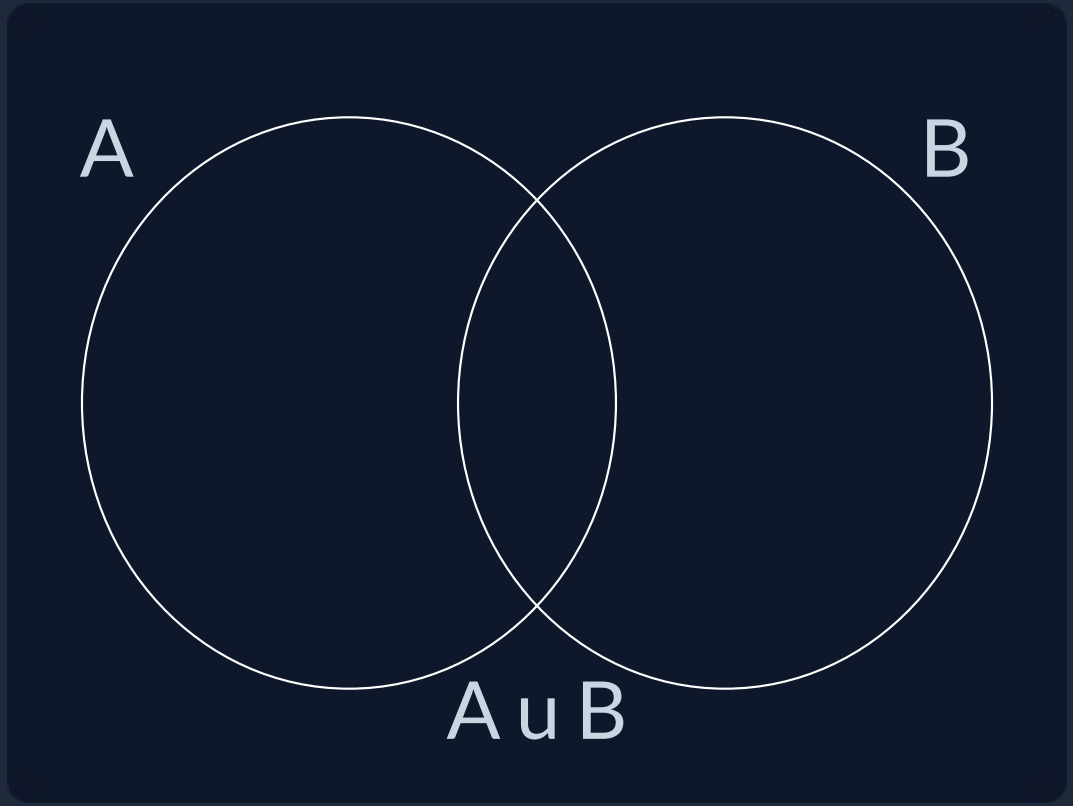
Probabilidad Condicionada

Definición: La probabilidad de que ocurra el evento A, sabiendo que el evento B ya ha ocurrido.

$$P(A \mid B)$$

Interpretación Práctica:

- Es la base de la actualización de información.
- Reduce el espacio muestral al evento B.
- Información parcial que refina nuestra estimación.



De Condicionada a Bayes

Partimos de la definición de probabilidad conjunta:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Por simetría, sabemos que $P(A \cap B) = P(B \cap A)$.

Esto nos permite "invertir" los condicionamientos.

El mensaje clave: Bayes nos permite ir de **P(Datos|Modelo)** a **P(Modelo|Datos)**.

Teorema de Bayes (Forma Estándar)

$$P(\theta|x) \propto P(x|\theta) \cdot P(\theta)$$

P(θ)

Prior: Lo que creíamos antes de ver los datos.

P(x|θ)

Verosimilitud: Qué tan probables son los datos según el modelo.

P(θ|x)

Posterior: Nuestra creencia actualizada.

Ejemplo Trabajado: Moneda Sesgada

El Escenario

- **Prior:** Creemos que la moneda es justa (pico en 0.5), pero no estamos 100% seguros.
- **Datos:** 10 lanzamientos, 8 caras (sesgo aparente hacia cara).
- **Posterior:** La curva se desplaza hacia la derecha (hacia 0.8), combinando la creencia inicial con la evidencia fuerte.

Lectura de "Razonamiento"

Pensamiento tipo "Diagnóstico"

Base

El **Prior** establece el punto de partida. Sin evidencia extraordinaria, nos mantenemos cerca de la base.

Peso de la Evidencia

La **Verosimilitud** actúa como el peso. Si los datos son muy improbables bajo el modelo actual, fuerzan un cambio drástico en la creencia.

"Afirmaciones extraordinarias requieren evidencia extraordinaria."

Distribución A Priori (Prior)

Informativa

Incorpora conocimiento experto fuerte o estudios previos.

Ej: "Sabemos que la tasa de conversión ronda el 2%."

Débilmente Informativa / Plana

Deja que los datos hablen por sí mismos.

Supuestos conservadores.

Ej: "La tasa está entre 0 y 1, todas igual de probables."

Fuentes: Expertos de dominio, literatura previa, lógica física.

Verosimilitud (Likelihood)

El puente entre los datos y el parámetro

- **Definición:** Es la probabilidad de observar los datos que tenemos, asumiendo un valor específico de θ.
- **Advertencia:** No es una distribución de probabilidad de θ. No suma a 1.
- **Intuición:** "¿Si el parámetro fuera X, qué tan raro sería ver estos datos?"

$$L(\theta \mid x) = P(x \mid \theta)$$

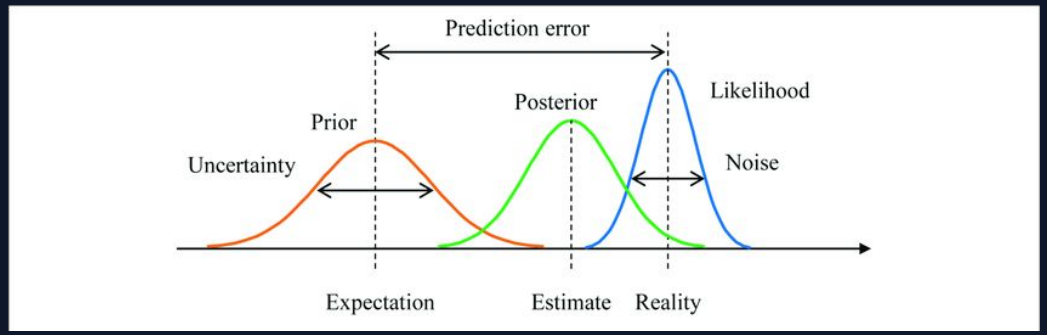
Posterior: El Resultado Central

Interpretación

Es la creencia actualizada. Contiene toda la información disponible (previa + datos) sobre el parámetro.

Regla de Oro

Posterior $\propto$ Likelihood $\times$ Prior



Resúmenes de la Posterior

Una distribución completa es útil, pero a veces necesitamos tomar una decisión puntual.

Media

Minimiza el error cuadrático medio.

Mediana

Minimiza el error absoluto medio (robusto).

Moda (MAP)

Máximo A Posteriori. El valor más probable.

Idea Clave: El "mejor" estimador depende del coste de equivocarse (Función de Pérdida).

Intervalos: Credibilidad vs Confianza

Bayesiano (Credibilidad)

"Hay un 95% de probabilidad de que el parámetro θ esté dentro de este rango."

Interpretación Directa e Intuitiva.

Frecuentista (Confianza)

"Si repitiéramos el experimento infinitas veces, el 95% de los intervalos calculados contendrían el parámetro real."

Interpretación sobre el procedimiento, no el parámetro.

Predicción Bayesiana

Objetivo DS: No solo estimar parámetros, sino predecir nuevos datos.

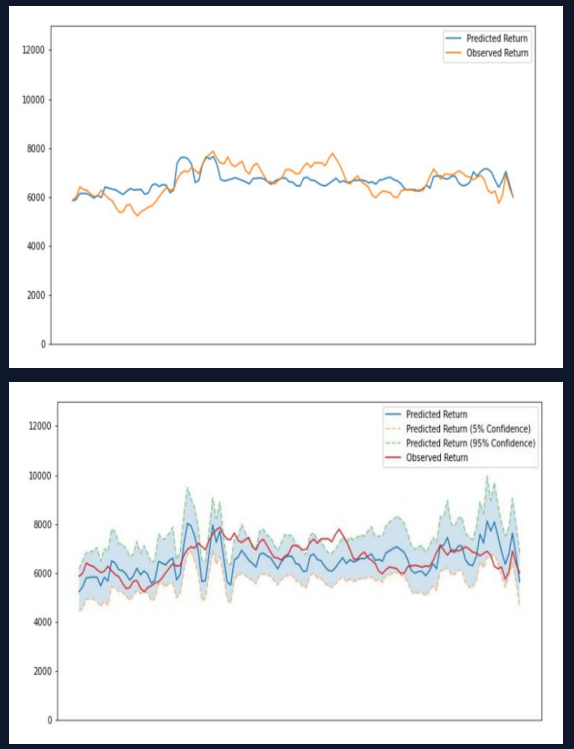
Posterior Predictiva

Calculamos la probabilidad de un nuevo dato (x')

integrando sobre todos los posibles valores de θ,

ponderados por su probabilidad posterior.

Predicción con incertidumbre incorporada.



Cómputo: El Reto de la Integración

El denominador de Bayes (Evidencia) requiere integrar sobre todos los parámetros posibles. A menudo, esto no tiene solución analítica cerrada.

Solución Numérica

Aproximación por simulación.

MCMC

Markov Chain Monte Carlo. Muestrear la posterior en lugar de resolverla.

Trade-off

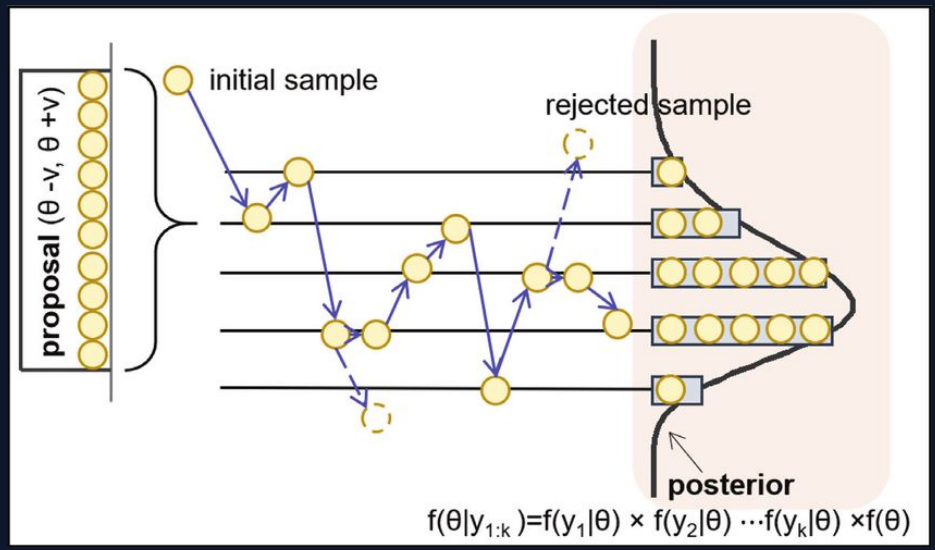
Flexibilidad del modelo vs Coste computacional.

Aclaración: Markov Chain Monte Carlo

• **Método Monte Carlo**: Es una técnica de aproximación numérica que utiliza la generación de muestras aleatorias para estimar funciones de probabilidad complejas. Permite obtener resultados mediante **simulación computacional**, donde cada ejecución puede arrojar datos ligeramente diferentes debido a su naturaleza estocástica.

• **Cadenas de Markov**: Representan un proceso donde la probabilidad de que ocurra un evento futuro depende **exclusivamente del estado actual** y no de la sucesión de eventos pasados. Esta característica, llamada **propiedad de Markov**, permite modelar sistemas con una "memoria" limitada únicamente al presente.

• **Markov Chain Monte Carlo (MCMC)**: Es una familia de algoritmos que combina la simulación aleatoria con las cadenas de Markov para muestrear distribuciones de probabilidad en espacios de alta dimensión. Su objetivo es construir una cadena cuya distribución final sea la distribución a posteriori deseada, facilitando la inferencia en modelos muy complejos.



Sensibilidad al Prior y Validación

- **Subjetividad Formalizada:** El prior debe ser transparente y justificado.
- **Análisis de Sensibilidad:** ¿Cambia mi conclusión si uso un prior diferente? Si cambia mucho, los datos no son suficientes.
- **Posterior Predictive Checks:** Simular datos desde el modelo ajustado y comparar con datos reales para validar.

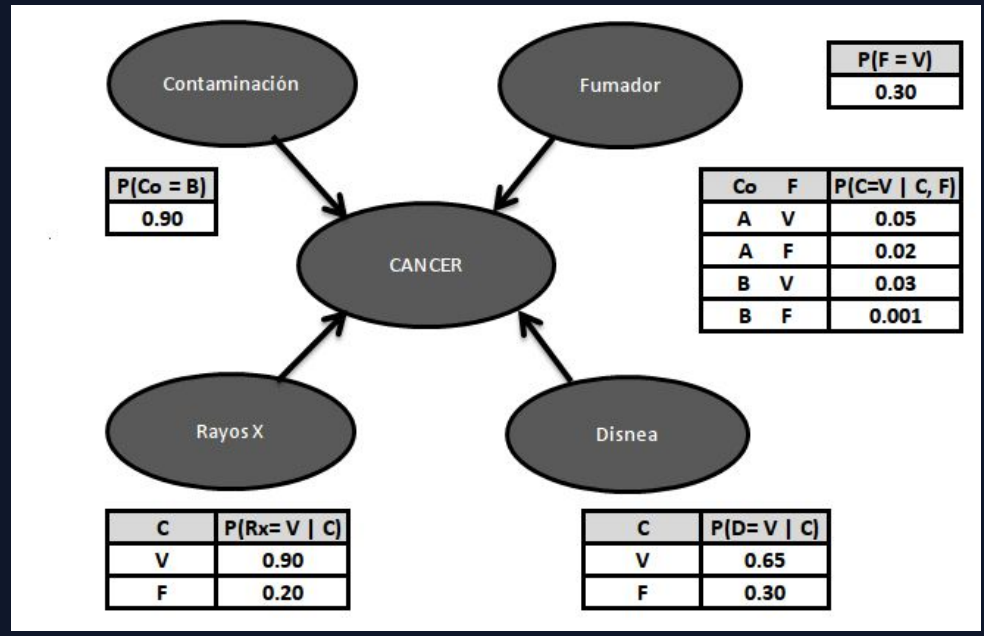
¿Qué es una Red Bayesiana?

Es un **Modelo Probabilístico Gráfico (DAG)**.

- **Nodos:** Variables aleatorias.
- **Aristas:** Dependencias condicionales (Causalidad).

Ventaja en DS: Permite manejar alta dimensionalidad factorizando el problema.

Interpretabilidad estructural.



Independencia Condicional y Factorización

El poder del grafo

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Padres}(X_i))$$

En lugar de una tabla gigante de probabilidad conjunta, descomponemos el problema en piezas locales pequeñas. Si A (X) no está conectado a B (Padres), son condicionalmente independientes.

Inferencia en Redes

Tipos de Consultas

- **Diagnóstico:** Observo el Efecto → Infiero la Causa.
- **Predicción:** Observo la Causa → Infiero el Efecto.

Algoritmos

- Propagación de Creencias.
- Eliminación de Variables.
- *Nota: Exacto es NP-Hard, se usan aproximaciones (MCMC) en redes grandes.*

Aprendizaje en Redes

¿Cómo construimos la red?

Aprendizaje de Estructura

Descubrir el grafo (aristas) a partir de los datos. (Score-based vs Constraint-based).

Aprendizaje de Parámetros

Dada la estructura, estimamos las tablas de probabilidad (CPTs) usando los datos.

Conexión

Es análogo al flujo "Prior + Datos" pero aplicado a la topología del grafo.

Clasificación Bayesiana: Regla de Decisión

Objetivo: Asignar una clase (y) a un dato (x).

$$\text{MAP: } \operatorname*{argmax}_{y} P(y \mid x)$$

Elegimos la clase que maximiza la probabilidad posterior. Esto está matemáticamente conectado con minimizar la función de pérdida 0-1 (error de clasificación).

Naive Bayes

El Supuesto "Ingenuo"

Asumimos que todas las características (features) son **condicionalmente independientes** dada la clase.

A pesar de ser un supuesto fuerte (y a menudo falso), funciona sorprendentemente bien en la práctica.

Ventajas

- Extremadamente rápido.
- Muy robusto en alta dimensionalidad (ej. Texto/NLP).
- Requiere pocos datos para entrenar.

Estimación y Suavizado

El problema de los ceros: Si una palabra no aparece en el training set para una clase, su probabilidad es 0, anulando toda la ecuación.

Suavizado de Laplace

Sumar 1 a los conteos. Equivale a poner un Prior Uniforme (Dirichlet) débil.

Calibración

Naive Bayes da buenas clasificaciones (ranking) pero probabilidades extremas (mal calibradas).

Cierre: ¿Cuándo usar Bayes en DS?

laSalle
UNIVERSITAT RAMON LLULL

Checklist de Selección

- Datos escasos (Small Data).
- Existe conocimiento previo fuerte.
- Necesidad crítica de medir incertidumbre (intervalos).
- Modelos Jerárquicos.

Takeaway

"La incertidumbre no es un error, es información."

Ejemplo – Vacuna del COVID

Nos encontramos con una nueva variante de COVID con una tasa de infección relativamente baja (0,75%) pero con una tasa de mortalidad relativamente alta.

Contamos con unos test rápidos que tienen los siguientes resultados:

- 98% de aciertos
- 1.5% de resultados inconcluyentes
- 0.5% de errores

PREGUNTA:

¿Cuántos test tenemos que hacernos para identificar una persona con la nueva variante de COVID para estar seguros con un 95% de certeza?

Ejemplo – Vacuna del COVID

PREGUNTA:

¿Cuántos test tenemos que hacernos para identificar una persona con la nueva variante de COVID para estar seguros con un 95% de certeza?

Prevalencia: $\pi = P(\text{infectado}) = 0.0075$

Cada test (independiente) da:

- **Correcto** 98% ⇒ si estás infectado sale "+", si no lo estás sale "−".
- **Error** 0.5% ⇒ invierte el signo (falso + o falso −).
- **Inconcluyente** 1.5% ⇒ no aporta información (y ocurre igual estés o no infectado).

Conocimiento adquirido:

$P(+ \mid \text{infectado}) = 0.98$ $P(+ \mid \text{no infectado}) = 0.005$

Ejemplo – Vacuna del COVID

PREGUNTA:

¿Cuántos test tenemos que hacernos para identificar una persona con la nueva variante de COVID para estar seguros con un 95% de certeza?

1) ¿Cuánta "certeza" da un solo positivo?

$$P(inf|+) = (\pi \cdot 0.98) / (\pi \cdot 0.98 + (1-\pi) \cdot 0.05)$$

Sustituyendo $\pi=0.0075$:

$$P(inf|+) \approx 0.597$$

Un único positivo te deja en ~59.7%, lejos del 95%.

2) Dos positivos independientes

$$P(inf|+ +) = (\pi \cdot 0.98^2) / (\pi \cdot 0.98^2 + (1-\pi) \cdot 0.05^2)$$

$$P(inf \mid ++) \approx 0.9966$$

Para superar el **95% de certeza**, necesitas **2 tests con resultado positivo concluyente** (dos "+"). Con dos positivos la certeza es **~99.66%** bajo estas hipótesis.

Anexos: Matriz de Confusión

Métricas asociadas Curva a la matriz de confusion binaria:

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		