

CRISTIAN SDROJEK

Notas de Distribuciones de probabilidad



Con un enfoque transversal, intuitivo y geométrico

Nota previa

Este texto está construido como un pseudodiálogo. Por un lado, mis preguntas —en el caso de los ejercicios, mis resoluciones parciales a los mismos—, y, por el otro, las respuestas que la IA —particularmente Gemini 3 Pro— ofrece a ellas. Es especialmente importante dejar constancia de los errores para regresar a ellos en el futuro, y recién ahora se hace posible mecanizar sistemáticamente este proceso de manera eficiente vía la IA. Es este un documento fiel a ese proceso de construcción de las ideas, mostrando los aciertos, pero también los más ridículos errores (debidamente señalados). Unos días pude ser más creativo, otros días estaba más despistado o fiel al curso académico oficial.

La idea de “apuntes” me parece una realidad profundamente obsoleta, así que opto por desarrollar este método, que incide en los aspectos que esos no pueden abarcar, como las tan numerosas y necesarias digresiones que van aflorando. Como es natural, la excelencia de cada uno de nosotros en ejecutarlo consiste en hacer las preguntas correctas, y dar con la tecla de por dónde pueden ir los tiros a la hora de responderlas, en lo cual consiste, precisamente, la ciencia, todo lo opuesto a aprender mecánicamente métodos desarrollados por otros —lo cual no significa que las reflexiones no se apoyen precisamente sobre esos métodos y descubrimientos—. Habida cuenta de que el tiempo requerido para así hacerlo es realmente elevado, quizás pueda ser de utilidad a alguien más la resolución a mis dudas personales, colocadas en un orden concreto y destiladas las preguntas y respuestas de la paja y fricción que orgánicamente surgen en el proceso de aprendizaje y elaboración de unas notas.

Índice

Nota previa	ii
1 Introducción a distribuciones unidimensionales	1
1.1 Distribución normal	1
1.1.1 Demostración de Herschel-Maxwell	1
1.1.2 Historia y enfoques	3
1.1.3 La genialidad de Gauss	4
1.2 Proceso de Poisson	6
1.2.1 Estocástica	6
1.2.2 De Poisson a las distribuciones	7
1.2.3 Binomial, Geométrica, Poisson y Exponencial	9
1.3 Distribuciones elementales y no elementales	11
1.4 Distribución binomial negativa	14
1.4.1 Memoria	14
1.4.2 Binomio generalizado de Newton (combinatoria)	14
1.5 Distribución hipergeométrica	16
1.5.1 Visualización gráfica	18
1.5.2 La binomial como límite de la hipergeométrica	20
1.5.3 Condicionamiento e hipersuperficie	22
1.5.4 Identidad de Vandermonde: dos visiones alternativas	26
1.6 Memoria	28
1.7 Distribución Gamma	29
1.7.1 Gamma como suma natural de exponenciales: visión del simplex	29
1.7.2 Visualización de distribución Gamma	33
1.7.3 Convolución, binomial negativa e hipergeométrica	35
1.7.4 Gamma como límite de binomial negativa	40
1.7.5 Función Gamma, recursividad y fórmula de Stirling	42
1.7.6 Generalización a menores de 1	47
1.7.7 Gamma como normal al cuadrado	51

1.8	Distribución Beta	54
1.8.1	Introducción	54
1.8.2	Otra forma de obtener la Beta	58
1.8.3	Distribución polar probabilística	61
1.8.4	Deducción de Beta a partir de Gamma	62
1.8.5	Visualización gráfica: convexidad, colas e inercia	65
1.9	Intuición tras el Teorema Central del Límite	79
1.9.1	Dos ejemplos	80
1.9.2	Señal, ruido y coeficiente de variación	81
1.9.3	¿Cuándo falla?	82
1.9.4	Dependencia e independencia de los errores	84
1.9.5	El TCL y la hipergeométrica	86
1.10	Distribución de Cauchy	87
1.11	Esquema general y deducción ultra rápida de las distribuciones	90
1.11.1	Deducción rápida de Beta a partir de Gamma	92
1.11.2	Deducción rápida de la FGM de Gamma	94
2	Distribución multinomial	95
2.1	Intuición	95
2.2	Vector aleatorio multinomial	99
2.2.1	Definición	99
2.2.2	Teorema (función de probabilidad) *	99
2.2.3	Componente redundante y binomio de Newton generalizado	99
2.3	El Teorema Multinomial	100
2.3.1	Teorema (Teorema Multinomial)	101
2.4	Relación con la distribución binomial	101
2.4.1	Caso $k = 2$ *	101
2.5	Propiedades del vector multinomial	102
2.5.1	Función generatriz de momentos *	102
2.5.2	Distribución marginal de las componentes *	102
2.5.3	Agrupación de componentes *	103
2.5.4	Momentos *	104
2.5.5	Reproductividad *	105
2.5.6	Distribución condicionada *	105
2.5.7	Curva de regresión *	106
3	Distribución multivariante	107
3.1	Vector normal multivariante	107
3.1.1	Definición	107
3.2	Propiedades del vector normal multivariante	107
3.2.1	Función generatriz de momentos *	107
3.2.2	Media y matriz de covarianzas *	108
3.2.3	Caso no singular: función de densidad *	108
3.2.4	Caso bidimensional *	109
3.2.5	Caso singular *	110
3.2.6	Distribución de los subvectores *	110
3.2.7	Incorrelación implica independencia *	111

3.2.8	Distribuciones condicionadas *	111
3.2.9	Curvas de regresión *	111
3.3	Intuición	112
3.3.1	La caja contenedora: la forma exacta de la media y la varianza condicionadas	118
3.3.2	El milagro lineal: la curva de regresión $h^*(x)$ en tres universos	121
3.4	¿Por qué la multinomial y la normal multivariante?	123
3.5	Ejemplos	126

Introducción a distribuciones unidimensionales

1.1 Distribución normal

Esta demostración es la de Herschel-Maxwell (1850) y es la única que deduce la fórmula exacta partiendo de la nada. Aquí tienes la deducción paso a paso, sin saltos, para que veas por qué *tiene* que ser esa fórmula y no otra.

Paso 1: las dos condiciones (el modelo)

Estamos lanzando dardos al centro $(0, 0)$. Definimos $f(x)$ como la probabilidad de caer en una coordenada x (error horizontal) y $f(y)$ para la vertical.

Condición A (independencia): asumimos que equivocarse en horizontal no afecta a equivocarse en vertical. La probabilidad conjunta de caer en el punto (x, y) es el producto:

$$P(x, y) = f(x) \cdot f(y)$$

Condición B (simetría rotacional): asumimos que la diana no tiene “preferencia” por ningún ángulo; la probabilidad solo depende de la distancia al centro (r) . Por Pitágoras, $r^2 = x^2 + y^2$:

$$P(x, y) = g(r) = g(\sqrt{x^2 + y^2})$$

Paso 2: el puzle matemático

Igualemos las dos condiciones. Buscamos una función f desconocida que cumpla:

$$f(x) \cdot f(y) = g(\sqrt{x^2 + y^2})$$

A la izquierda, una multiplicación de variables separadas; a la derecha, una función que depende de la suma de cuadrados ($x^2 + y^2$). ¿Cómo convertimos una multiplicación en una suma? Solo hay una herramienta en matemáticas que hace eso: el logaritmo.

Paso 3: la deducción

Aplicamos logaritmo natural a ambos lados:

$$\ln(f(x) \cdot f(y)) = \ln(g(r))$$

Por la propiedad $\ln(a \cdot b) = \ln a + \ln b$:

$$\ln(f(x)) + \ln(f(y)) = \ln(g(r))$$

Llamemos $h(x) = \ln(f(x))$. La ecuación queda:

$$h(x) + h(y) = h(\sqrt{x^2 + y^2})$$

Buscamos una función tal que $h(x) + h(y)$ dependa solo de $x^2 + y^2$. La única función simple que lo cumple es la cuadrática. Si $h(x) = C \cdot x^2$:

$$C \cdot x^2 + C \cdot y^2 = C(x^2 + y^2) = C \cdot r^2$$

La ecuación cierra perfectamente. Esto implica que el logaritmo de la probabilidad tiene que ser algo por x^2 :

$$\ln(f(x)) = C \cdot x^2$$

Paso 4: despejando la fórmula final

Deshacemos el cambio para encontrar la $f(x)$ original. De $\ln(f(x)) = C \cdot x^2$, quitamos el logaritmo con la exponencial:

$$f(x) = e^{C \cdot x^2}$$

Si C fuera positivo (p. ej. e^{x^2}), la probabilidad crecería al infinito cuanto más lejos cayera el dardo: imposible. Por tanto, C tiene que ser negativo; llamémoslo $-k$:

$$f(x) = e^{-k \cdot x^2}$$

Resumen de tus dudas específicas

¿Por qué e ? Porque para resolver la ecuación $f(x) \cdot f(y) = g(r)$ tuvimos que usar logaritmos (para convertir producto en suma). La inversa del logaritmo es e : es la única base que conecta producto con suma de forma natural.

¿Por qué el cuadrado (x^2)? Porque la distancia en el plano es pitagórica ($r^2 = x^2 + y^2$). La única forma de que las partes (x, y) sumen para dar el todo (r) de forma simétrica es que la función dependa del cuadrado.

¿Por qué está partido por 2? Al final nos queda e^{-kx^2} , donde k es un número libre que controla la anchura. Los matemáticos definieron la desviación típica σ tal que $\sigma^2 = \frac{1}{2k}$, para que al derivar más adelante el 2 del exponente baje ($2x$) y cancele al 2 del denominador. Es pura comodidad estética.

Conclusión: la Normal no se “inventó” ni se eligió por capricho. Es la única solución matemática posible al problema: una distribución de errores independiente en coordenadas y simétrica en rotación.

La Normal es una entidad que “vive” cómodamente en 2D (o n D) gracias a su simetría rotacional, y lo que vemos en 1D es solo su proyección o corte. Tanto para deducir su fórmula (argumento de los dardos) como para calcular su área (integral de Gauss), el truco es el mismo: subir una dimensión para explotar la simetría circular y luego bajar. Es como si, para entender la sombra de una esfera, tuvieras que mirar la esfera completa.

Respecto a la historia: no fue con los dardos. Esa demostración tan elegante (Herschel-Maxwell) llegó mucho más tarde (1850–1860). El origen es una “guerra” entre dos titanes y dos enfoques distintos.

1. **De Moivre (1733), el vago genial.** Abraham de Moivre se ganaba la vida calculando probabilidades para apostadores en Londres. Su problema era calcular una Binomial con n gigante: la probabilidad de sacar 500 caras en 1000 monedas exige un $1000!$, imposible de calcular a mano. Su solución fue usar aproximaciones (la fórmula de Stirling para los factoriales), y descubrió que, entre todo ese lío, emergía una curva suave con la forma e^{-x^2} . Para él no era una “ley universal”, sino solo un truco de cálculo: un enfoque teórico-analítico, no empírico ni geométrico.
2. **Gauss (1809), la ingeniería inversa.** Aquí nace el nombre “Gaussiana”. Gauss estaba calculando la órbita del asteroide Ceres con muchas mediciones sujetas a error, e hizo un razonamiento filosófico brutal (inverso al de los dardos). Axioma: si mido varias veces una cosa, la mejor estimación de la verdad es la media aritmética. Pregunta inversa: ¿cómo tiene que ser la distribución de los errores para que la media aritmética sea el estimador de máxima verosimilitud? Gauss demostró que la única función de error que hace óptimo al promedio es la forma e^{-x^2} . No dijo “los errores parecen seguir esta curva”, sino: si crees que hacer promedios funciona, estás asumiendo implícitamente que los errores son Normales.
3. **Laplace (1810), el caos genera orden.** Laplace conectó lo de Gauss con lo de De Moivre y demostró el Teorema Central del Límite. Su aporte: no importa si el error viene de un dado, de una moneda o de un dardo; si sumas muchas causas pequeñas, siempre sale la

curva de Gauss. Aquí dejó de ser un “truco” o una “hipótesis” y se convirtió en una ley universal.

4. **Maxwell (1860), los dardos.** Mucho después, Maxwell estudiaba gases (moléculas chocando) y usó el argumento de los dardos (independencia de x, y, z más simetría rotacional de la velocidad) para deducir cómo se mueven las partículas. Es la demostración más bonita y geométrica, pero históricamente fue la última en llegar, cuando ya sabíamos que la fórmula funcionaba.

Resumen para tu mapa mental

- **De Moivre:** es un atajo para no calcular factoriales gigantes (Binomial $\rightarrow$ Normal).
- **Gauss:** es la consecuencia lógica de usar el promedio (error $\rightarrow$ media).
- **Maxwell:** es la única forma de tener simetría en el espacio (geometría 2D $\rightarrow$ 1D).

La intuición de “extender a $\mathbb{R}^2$ y volver” es la visión moderna y geométrica (Maxwell), y es la mejor para visualizar por qué aparecen π y e .

La genialidad de Gauss (1809) es que se atrevió a hacer ingeniería inversa a la realidad. Mientras que Herschel (el de los dardos) miró al mundo (la simetría) y dedujo la fórmula, Gauss miró a los humanos (cómo calculamos cosas) y dedujo cómo tenía que ser el mundo para que tuviéramos razón. Aquí tienes el razonamiento paso a paso; es una belleza lógica.

1. **El axioma “sagrado” (la trampa).** Gauss partió de una verdad aceptada por todos los astrónomos de la época: si mides la posición de una estrella 10 veces y obtienes 10 números distintos, la mejor estimación de la verdad es la media aritmética. Nadie lo demostraba; se usaba porque funcionaba y era intuitivo. Gauss dijo: aceptemos que el promedio es la respuesta correcta; si lo es, ¿cómo tiene que ser la forma de los errores?
2. **La intuición de la máxima verosimilitud.** Imagina que miras una estrella con tu telescopio para saber dónde está exactamente (su coordenada X). Digamos que la verdad absoluta es que la estrella está en la posición 100. Haces 3 mediciones, pero como el telescopio no es perfecto, la atmósfera tiembla y tu pulso falla, mides $x_1 = 101$, $x_2 = 99$, $x_3 = 103$. El error es la distancia entre lo medido y la verdad: $+1$, -1 y $+3$.

El problema: tú tienes los papeles con el 101, el 99 y el 103, pero no sabes que la verdad es 100; quieres descubrir ese 100. Llamemos μ a esa verdad desconocida. Tus errores son:

$$e_1 = 101 - \mu, \quad e_2 = 99 - \mu, \quad e_3 = 103 - \mu$$

Gauss se pregunta: ¿cuál debería ser el valor de μ para que haber obtenido esos tres números sea lo más lógico posible? Por reducción al absurdo: si $\mu = 1000$, para medir 101 habrías cometido un error de -899 , algo bajísimo en probabilidad; muy poco verosímil. Si $\mu = 101$, los errores serían 0 , -2 y $+2$: pequeños, bastante verosímil. Gauss busca el valor exacto de μ que hace que el conjunto de errores sea el más probable posible.

3. **La ingeniería inversa de Gauss.** La función incógnita: no sabemos qué forma tiene la curva de probabilidad de los errores $f(e)$; solo sabemos que los errores pequeños son más probables que los grandes. La probabilidad de que ocurran las tres mediciones a la vez es el producto de sus probabilidades individuales:

$$P = f(101 - \mu) \cdot f(99 - \mu) \cdot f(103 - \mu)$$

El axioma del promedio dice que la mejor estimación de μ es la media de las mediciones, $(101 + 99 + 103)/3 = 101$. Gauss obliga entonces a las matemáticas a cumplir esto: busco una función f tal que, cuando maximice la probabilidad P , el resultado sea exactamente el promedio.

4. La deducción.

1. Planteamiento. Tenemos n mediciones independientes $M_1, \dots, M_n$, un valor verdadero desconocido μ , y errores $x_i = M_i - \mu$. La función de probabilidad del error es $\phi(x)$ (desconocida, es lo que buscamos). La verosimilitud L es el producto:

$$L = \phi(M_1 - \mu) \cdot \phi(M_2 - \mu) \cdots \phi(M_n - \mu)$$

2. Maximizar la probabilidad. Para hallar el máximo de L respecto a μ , derivamos e igualamos a cero. Maximizar el producto es un infierno, pero maximizar su logaritmo es fácil (y el máximo está en el mismo sitio porque el logaritmo es creciente):

$$\ln L = \ln(\phi(M_1 - \mu)) + \cdots + \ln(\phi(M_n - \mu))$$

Derivamos respecto a μ (regla de la cadena, la derivada de lo de dentro es -1):

$$\frac{d(\ln L)}{d\mu} = \frac{\phi'(M_1 - \mu)}{\phi(M_1 - \mu)}(-1) + \frac{\phi'(M_2 - \mu)}{\phi(M_2 - \mu)}(-1) + \cdots = 0$$

Definimos la “derivada logarítmica” auxiliar $\psi(x) = \frac{\phi'(x)}{\phi(x)}$. La condición de máximo es:

$$-\sum_{i=1}^n \psi(M_i - \mu) = 0 \implies \sum_{i=1}^n \psi(M_i - \mu) = 0$$

3. Imponer el axioma del promedio. Esta ecuación debe ser cierta si y solo si μ es el promedio aritmético. Si $\mu = \bar{M}$, por definición de promedio la suma de las desviaciones es cero:

$$\sum_{i=1}^n (M_i - \bar{M}) = 0$$

Gauss tiene entonces dos ecuaciones que deben cumplirse a la vez para cualquier conjunto de datos, con $x_i = M_i - \bar{M}$: $\sum \psi(x_i) = 0$ (máxima verosimilitud) y $\sum x_i = 0$ (propiedad del promedio).

4. Resolver la ecuación funcional (el paso crítico). Buscamos la forma de $\psi(x)$ tal que, siempre que la suma de los números sea 0, la suma de las funciones también sea 0:

$$\sum x_i = 0 \implies \sum \psi(x_i) = 0$$

Con $n = 2$: si $x_1 + x_2 = 0$, entonces $x_2 = -x_1$, y la condición da $\psi(x_1) + \psi(-x_1) = 0$, es decir, $\psi(x) = -\psi(-x)$: ψ es impar. Con $n = 3$: sean x_1, x_2 cualesquiera y $x_3 = -(x_1 + x_2)$; la condición dice $\psi(x_1) + \psi(x_2) + \psi(-(x_1 + x_2)) = 0$, y usando que ψ es impar:

$$\psi(x_1) + \psi(x_2) = \psi(x_1 + x_2)$$

Esta es la famosa ecuación funcional de Cauchy: la función de la suma es la suma de las funciones. Su única solución continua es una recta por el origen:

$$\psi(x) = k \cdot x$$

5. Integrar para volver a ϕ . Como $\psi = \phi' / \phi$, tenemos $\frac{\phi'(x)}{\phi(x)} = kx$, una ecuación diferencial de variables separables:

$$\frac{d\phi}{\phi} = kx \, dx \implies \ln(\phi(x)) = \frac{kx^2}{2} + C$$

Despejamos ϕ elevando a e :

$$\phi(x) = e^{\frac{kx^2}{2} + C} = e^C \cdot e^{\frac{k}{2}x^2}$$

Con $A = e^C$ (amplitud), y como la curva debe bajar (no subir al infinito) para ser probabilidad, el exponente debe ser negativo; hacemos $\frac{k}{2} = -h^2$. Resultado final:

$$\phi(x) = Ae^{-h^2x^2}$$

5. Resumen de la rigurosidad. La máxima verosimilitud nos da una suma de derivadas logarítmicas igual a 0; el axioma del promedio nos dice que la suma de los argumentos es 0; la ecuación de Cauchy obliga a que la relación entre ambas sumas sea lineal ($\psi(x) = kx$); y la ecuación diferencial convierte esa linealidad en un exponente cuadrático (x^2). Así se demuestra, sin dardos ni magia, que si crees en el promedio, matemáticamente tienes que creer en la Normal.

1.2 Proceso de Poisson

Si la Probabilidad fuera la Física, la “estocástica” sería la cinemática y la dinámica. Hasta ahora hemos hablado de “fotos” estáticas: una variable X (Normal, Binomial...) es una foto fija — tiras el dado, sale un 5, fin de la historia. La estocástica (o procesos estocásticos) es cuando conviertes esa foto en una *película*: introduces el tiempo t . Ya no es X , ahora es $X(t)$.

Un proceso estocástico es una variable aleatoria que evoluciona y cambia a medida que pasa el tiempo.

1. La diferencia visual: el borracho y la farola

Usamos el modelo más famoso de la historia: el camino aleatorio (*random walk*).

Escenario estático (lo que hemos visto): tienes a un borracho en un bar. Le preguntas: “si das un paso, ¿dónde acabas?”. Respuesta: 50% un metro adelante, 50% un metro atrás. Esto es una variable Bernoulli (o Rademacher): una sola acción.

Escenario dinámico (lo nuevo): el borracho sale del bar a las 00:00 y da un paso aleatorio cada segundo. ¿Dónde estará a las 00:15? Esto ya no es tirar una moneda: es una *trayectoria*.

$$X(t) = X(t - 1) + \text{paso aleatorio}$$

Para saber dónde está en $t = 100$ necesitas saber dónde estaba en $t = 99$: hay dependencia temporal. La historia importa (o al menos el estado actual importa).

2. Los dos sabores del tiempo

Igual que las variables eran discretas o continuas, ahora el tiempo también puede serlo. Hay 4 combinaciones, pero las 2 reinas son:

A. Tiempo discreto (cadenas de Markov): el tiempo avanza a saltos ($t = 1, 2, 3, \dots$), como turnos en un juego de mesa (p. ej. el precio de una acción al cierre de bolsa cada día). Concepto clave: la propiedad de Markov, “el futuro solo depende del presente, no del pasado”. Si el borracho está en la farola 5, da igual si vino desde la 4 o desde la 6: solo importa que está en la 5 ahora.

B. Tiempo continuo (procesos de difusión): el tiempo fluye suavemente ($t \in [0, \infty)$), como el movimiento de una partícula de polen en agua (movimiento browniano). Es el límite matemático cuando los pasos del borracho se hacen infinitesimalmente pequeños y rápidos.

3. La conexión física: difusión y entropía

Imagina 1.000 borrachos saliendo del mismo bar (origen $x = 0$) a la vez. En $t = 0$ todos están en el centro (varianza 0); en $t = 1$ unos avanzan y otros retroceden, empiezan a dispersarse; en $t = 1000$ unos están lejísimos a la derecha y otros a la izquierda. Si dibujas el histograma de las posiciones en cada instante, la distribución se va ensanchando: la media sigue siendo 0 (por simetría el centro de masas no se mueve), pero la varianza crece con el tiempo ($\sigma^2 \propto t$).

¿Qué forma tiene esa distribución que se ensancha? Por el Teorema Central del Límite (sumar muchos pasos), una campana de Gauss que se aplasta con el tiempo. Esto es, literalmente, la ecuación del calor o de difusión:

$$\frac{\partial f}{\partial t} = D \frac{\partial^2 f}{\partial x^2}$$

(la variación en el tiempo es proporcional a la curvatura en el espacio). La estocástica es la disciplina que estudia cómo la incertidumbre (entropía) se propaga a través del tiempo.

El Proceso de Poisson no es una “distribución” más: es la madre de todas las distribuciones de conteo y de tiempo, el modelo matemático de la aleatoriedad pura y continua. Si lo entiendes, entiendes la mitad de la asignatura de golpe. Olvida las fórmulas un segundo y piensa en lluvia cósmica.

1. ¿Qué es el Proceso de Poisson? (la película frente a la foto)

La mayoría de distribuciones son “fotos” estáticas; el Proceso de Poisson es una *película*: describe eventos que ocurren a lo largo del tiempo (o del espacio) de forma totalmente descorrelacionada. La intuición: imagina una línea de tiempo infinita y vas tirando “puntos” (eventos) sobre ella sin apuntar, con total indiferencia. No intentas espaciarlos (sería demasiado regular) ni agruparlos (sería contagio); simplemente dejas caer arena sobre una mesa.

2. Los tres supuestos fundamentales

Para que algo sea un Proceso de Poisson debe cumplir tres reglas. Si falla una, no es Poisson.

1. **Independencia (sin memoria):** lo que pase en el intervalo $[t_1, t_2]$ no tiene nada que ver con lo que pasó antes ni con lo que pasará después. Que haya caído un rayo hace 1 segundo no hace más ni menos probable que caiga otro ahora. Un terremoto (cuyas réplicas dependen del evento principal) *no* es Poisson.
2. **Estacionariedad (λ constante):** la tasa media de eventos por unidad de tiempo λ es constante; la “intensidad” de la lluvia no cambia. El tráfico en hora punta la viola (la λ a las 8:00 es distinta que a las 3:00): sería un Poisson no homogéneo.
3. **No simultaneidad:** la probabilidad de que ocurran dos eventos exactamente en el mismo instante infinitesimal dt es cero. Los eventos llegan en fila india, nunca dos a la vez.

3. El árbol genealógico: ¿quién nace de Poisson?

El Proceso de Poisson es el tronco; según qué le preguntes, obtienes una distribución distinta. Imagina que ves pasar coches (eventos) en una carretera desierta (Proceso de Poisson con tasa λ).

1. **¿Cuántos pasan en 1 hora?** → **Poisson**. Fijas la ventana de tiempo t y cuentas los eventos k :

$$P(N(t) = k) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$$

Es discreta: la “foto” de conteo de la película.

2. **¿Cuánto tiempo pasa entre un coche y el siguiente?** → **Exponencial**. Fijas el contador (quiero el siguiente evento) y mides el tiempo x :

$$f(x) = \lambda e^{-\lambda x}$$

Es continua: el “espacio vacío” entre puntos. La Exponencial es la única continua sin memoria, coherente con el supuesto 1 de Poisson.

3. **¿Cuánto tardan en pasar 5 coches?** → **Gamma (Erlang)**. Es la suma de 5 tiempos exponenciales independientes:

$$T_k = X_1 + X_2 + \dots + X_k$$

Si sumas exponenciales (esperas individuales) obtienes una Gamma: la distribución del tiempo total de espera hasta el evento k -ésimo.

4. **Si sé que han pasado 10 coches en una hora, ¿dónde están?** → **Uniforme**. Esta es la más sorprendente. Si sabes que en el intervalo $[0, T]$ han ocurrido exactamente n eventos Poisson, esos n eventos están distribuidos uniformemente en el intervalo: condicionado al total, el desorden máximo temporal es la Uniforme. Si tiras los puntos al azar (Uniforme) y luego los ordenas, recuperas el proceso.
La posición de cada punto es Uniforme (indiferencia total), pero el patrón visual del conjunto es irregular (grumos y huecos): “Uniforme” se refiere a la igualdad de oportunidades de cada sitio para recibir un punto, no a la estética final del dibujo. La probabilidad de que un evento esté en el instante t es igual para todos los t ; si la representas frente al tiempo, sale una recta horizontal, que por definición es una Uniforme.
5. **¿Qué pasa si miro la carretera parpadeando cada segundo?** → **Bernoulli / Geométrica**. Si discretizas el tiempo en trocitos de tamaño Δt : en cada trocito, ¿pasa un coche? Sí/No → Bernoulli ($p \approx \lambda \Delta t$). ¿Cuántos trocitos vacíos hasta el primer coche? → Geométrica. ¿Cuántos trocitos con coche en total? → Binomial.

Resumen del árbol:

- **Poisson** (discreta): contar eventos en tiempo fijo.
- **Exponencial** (continua): medir tiempo hasta 1 evento.
- **Gamma** (continua): medir tiempo hasta k eventos.
- **Uniforme** (continua): ubicación de eventos condicionada al total.
- **Bernoulli/Geométrica**: versión pixelada de todo lo anterior.

El Proceso de Poisson es el “generador”; las distribuciones estáticas son distintas formas de cortar el salchichón de ese proceso: si cortas por tiempo → Poisson; si cortas por eventos → Exponencial/Gamma.

4. ¿Por qué es útil? (la ley de los números pequeños)

¿Por qué el mundo real se ajusta tanto a Poisson (llamadas telefónicas, desintegración atómica, errores en código, fotones llegando a un telescopio)? Por el Teorema de Superposición (Palm-Khintchine). Imagina un servidor web con 10.000 usuarios independientes, cada uno con su propio patrón raro y una probabilidad muy baja de entrar en este preciso milisegundo. El teorema dice: si sumas muchos procesos de renovación independientes (da igual sus distribuciones originales), la suma converge a un Proceso de Poisson. La naturaleza tiende a Poisson cuando agregas muchas causas independientes y raras: es el Teorema Central del Límite, pero para flujos de eventos en vez de para sumas de valores.

5. ¿Cómo se conecta con la entropía?

Si preguntas: “¿cuál es la forma de colocar puntos en una recta $[0, T]$ que maximice mi ignorancia sobre dónde están (máxima entropía), sabiendo solo la tasa media?”, la respuesta es el Proceso de Poisson. Cualquier otro proceso (que los puntos tiendan a juntarse, o a separarse regularmente) implica meter información (estructura). Poisson es el “silencio estructurado” del universo: el azar en su estado más puro.

1.2.3 Binomial, Geométrica, Poisson y Exponencial

1. Anotaciones rápidas de conexión

Binomial → **Poisson**: mantenemos la media $\mu = np$ constante (llamémosla λ en el intervalo)

mientras hacemos $n \rightarrow \infty$ y $p \rightarrow 0$. Físicamente, pasamos de contar éxitos en n ensayos discretos a contar eventos puntuales en un continuo.

Poisson $\rightarrow$ Exponencial: si $N(t)$ es un proceso de Poisson con tasa λ , la Exponencial es la distribución del “tiempo hasta el primer evento” T . Se deduce imponiendo que en el intervalo $(0, t]$ ocurran cero eventos ($N(t) = 0$).

2. Modo alternativo: la Exponencial como límite de la Geométrica

Discretizamos el tiempo continuo t en pequeños intervalos de tamaño Δt . Sea λ la tasa media de eventos por unidad de tiempo; la probabilidad de éxito en un intervalo infinitesimal Δt es $p = \lambda \Delta t$, y el número de intervalos (ensayos) necesarios para cubrir un tiempo t es $k = \frac{t}{\Delta t}$.

Función de supervivencia geométrica: la Geométrica mide el número de fracasos antes del éxito. La probabilidad de que el evento ocurra después del paso k (sobrevivir k pasos) es:

$$P(X > k) = (1 - p)^k$$

Paso al límite continuo: sustituimos p y k en función de t y hacemos $\Delta t \rightarrow 0$:

$$P(T > t) = \lim_{\Delta t \rightarrow 0} (1 - \lambda \Delta t)^{\frac{t}{\Delta t}}$$

Con el cambio $n = \frac{1}{\Delta t}$ (si $\Delta t \rightarrow 0$, entonces $n \rightarrow \infty$) aparece el límite de Euler:

$$P(T > t) = \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{nt} = \left[\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n\right]^t$$

Recordando que $\lim_{n \rightarrow \infty} (1 - x/n)^n = e^{-x}$:

$$P(T > t) = (e^{-\lambda})^t = e^{-\lambda t}$$

Hemos obtenido la función de supervivencia de la Exponencial.

3. Herencia de la falta de memoria

La Geométrica es la única discreta sin memoria. Al hacer el límite, la Exponencial hereda esta propiedad, convirtiéndose en la única continua sin memoria. En la Geométrica:

$$P(X > n + m \mid X > n) = \frac{(1 - p)^{n+m}}{(1 - p)^n} = (1 - p)^m = P(X > m)$$

En la Exponencial (aplicando el resultado anterior):

$$P(T > t + s \mid T > t) = \frac{P(T > t + s \text{ y } T > t)}{P(T > t)} = \frac{P(T > t + s)}{P(T > t)}$$

Sustituyendo $e^{-\lambda t}$:

$$= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = \frac{e^{-\lambda t} \cdot e^{-\lambda s}}{e^{-\lambda t}} = e^{-\lambda s} = P(T > s)$$

La probabilidad de esperar un tiempo adicional s es independiente de cuánto tiempo t ya hayas esperado: el sistema se “reinicia” instantáneamente en el continuo igual que lo hacía en el discreto.

4. Deducción de la función de densidad

Tenemos la probabilidad de que el evento ocurra después de t (supervivencia):

$$S(t) = P(T > t) = e^{-\lambda t}$$

La función de distribución acumulada es la probabilidad contraria (que ocurra antes o en t):

$$F(t) = P(T \leq t) = 1 - S(t) = 1 - e^{-\lambda t}$$

Y la densidad es la derivada de la acumulada respecto a t :

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = -(-\lambda)e^{-\lambda t} = \lambda e^{-\lambda t}$$

Y ahí tienes la definición estándar de la Exponencial.

1.3 Distribuciones elementales y no elementales

Usaremos esta distinción:

Distribuciones de máxima entropía (“elementales” o “naturales”): aparecen espontáneamente cuando *no* impones reglas mecánicas y solo conoces un dato estadístico (como el promedio). Son el estado de equilibrio del sistema.

Distribuciones mecanicistas (“constructivas”): aparecen porque has diseñado un experimento con reglas rígidas (un juego, un conteo fijo, un mazo de cartas). Tienen una “historia” detrás.

I. Las discretas (el mundo de los saltos)

1. **Bernoulli (el átomo)** — *elemental*. Es la unidad mínima de información probabilística; no se deriva de nada y todo lo demás se construye sobre ella. Si solo puedes distinguir “blanco/negro”, esta es la única matemática posible.

2. **Binomial (el acumulador)** — *mecanicista*. Es la suma de n Bernoullis independientes. No es libre, sino artificial, porque tú decides parar el juego exactamente en el intento n : es un contador de éxitos en un escenario controlado.
3. **Geométrica (la espera pura)** — *máxima entropía*. Es la distribución más honesta de los enteros positivos si solo conoces su media: la forma natural en que se distribuye la “mala suerte” (fracasos) hasta un éxito, sin asumir que el sistema se cansa o tiene memoria.
4. **Binomial negativa (la resistencia)** — *mecanicista*. Es la suma de r Geométricas. No modela el azar puro, sino la acumulación: “¿cuánto tardo en conseguir 5 éxitos?”. Al exigir r éxitos impones una estructura que el azar puro no tiene.
5. **Hipergeométrica (el mazo de cartas)** — *mecanicista con memoria fuerte*. Es una Binomial sin reposición: modela un sistema finito que se “agota” a medida que sacas cosas. La probabilidad cambia en cada paso porque la urna se hace más pequeña.
6. **Poisson (el flujo)** — *máxima entropía*. Es la distribución más honesta para contar eventos en un intervalo si solo conoces la tasa media y asumes independencia. Es el “ruido de fondo” del conteo.

II. Las continuas (el mundo del fluido)

7. **Uniforme (la indiferencia)** — *máxima entropía suprema*. Es la distribución más honesta en un intervalo acotado $[a, b]$ cuando no sabes absolutamente nada, ni siquiera la media. Asignar más probabilidad a un punto que a otro sería mentir.
8. **Exponencial (el olvido)** — *máxima entropía*. Es la distribución más honesta para una variable positiva (tiempo, distancia) si solo conoces su media. Es la única curva continua que no “envejece”: la aleatoriedad pura en el tiempo, sin memoria ni desgaste.
9. **Gamma (el desgaste)** — *mecanicista*. Es la suma de k Exponenciales; modela el tiempo necesario para que ocurran k eventos. A diferencia de la Exponencial, sí tiene memoria (sabe que ya han pasado $k - 1$ eventos), por lo que modela envejecimiento o acumulación de daño.
10. **Normal / Gaussiana (el equilibrio)** — *máxima entropía*. Es la distribución más honesta en toda la recta real si conoces la media y la varianza. Es el resultado inevitable de sumar muchas causas independientes (Teorema Central del Límite): el caos, dejado a su aire, siempre forma una campana.
11. **Beta (la plastilina)** — *funcional / herramienta*. No es un mecanismo físico simple ni una maximización de entropía básica, sino un comodín acotado en $[0, 1]$. Se usa por flexible: cambiando sus parámetros imita una Uniforme, una Normal o una Exponencial dentro de un rango fijo. Es la favorita para modelar porcentajes o probabilidades inciertas.

Jerarquía de restricciones

Podemos clasificar las distribuciones según el **precio de entrada** (la información mínima, o *estadísticos suficientes*) que necesitas entregarle al sistema para que te devuelva esa distribución.

Nivel 0: información cero. Solo conoces el escenario (los límites); no sabes nada del comportamiento.

Uniforme continua (a, b) :

- **Lo que necesitas:** “esto pasa entre a y b ”.
- **Asunciones:** ninguna. Ni siquiera necesitas la media: si la supieras, la Uniforme dejaría de ser la respuesta más honesta. Es la ignorancia perfecta en un recinto cerrado.

Nivel 1: solo escala (media/intensidad). Conoces el escenario más la escala.

Exponencial (λ):

- **Lo que necesitas:** “son valores positivos ($x > 0$) y el promedio es μ ”.
- **Asunciones:** la varianza no es un dato de entrada, es una consecuencia ($\sigma^2 = \mu^2$). El sistema es libre de fluctuar, solo tiene que respetar el promedio a la larga.

Geométrica (p):

- **Lo que necesitas:** “son enteros positivos y el promedio de intentos es μ ”.
- **Asunciones:** igual que la Exponencial; la varianza sale sola.

Poisson (λ):

- **Lo que necesitas:** “cuento cosas en un intervalo y el promedio es λ ”.
- **Asunciones:** la varianza es igual a la media por obligación ($\sigma^2 = \lambda$); no puedes elegirla.

Todas estas son de Nivel 1 porque tienen un solo parámetro: fijas la intensidad/media y la forma de la curva queda sellada. No existe “una Exponencial con media 5 pero muy ancha”: si la media es 5, el ancho es fijo.

Nivel 2: media más varianza. Conoces el escenario, la escala y la dispersión.

Normal (μ, σ^2):

- **Lo que necesitas:** “vive en la recta real, con centro en μ y energía σ^2 ”.
- **Por qué necesitas la varianza:** en la recta infinita, si solo fijas la media, la entropía intenta hacer la distribución infinitamente ancha (plana) para ser lo más uniforme posible; la “tarta” de probabilidad se estira tanto que la altura se hace cero y la distribución desaparece. Necesitas σ^2 como correa para que no se disperse hasta el infinito.

Diferencia clave Nivel 1 frente a Nivel 2: en la Exponencial el suelo ($x = 0$) actúa de tope, por eso no se escapa a $-\infty$ y no necesitas fijar la varianza; en la Normal no hay suelo, así que necesitas la varianza para atarla.

Nivel máximo: mecanicistas (no elementales). Conoces el escenario, los parámetros estadísticos y además las reglas del juego. Aquí la información necesaria no es solo estadística, es **estructural**: hay que definir la arquitectura del experimento.

Binomial (n, p):

- **Lo que necesitas:** la probabilidad p y además el número de intentos n .
- **El extra:** ese n no es una propiedad estadística (como la varianza), es una regla de parada: “el juego se acaba a la décima”.

Gamma (k, λ):

- **Lo que necesitas:** la tasa λ y además el número de eventos a esperar k .
- **El extra:** k es estructural; fuerzas al sistema a acumular memoria de k pasos.

Hipergeométrica (N, M, n):

- **Lo que necesitas:** el inventario completo de la urna (N, M) y la regla de extracción (n). Necesitas el plano completo del almacén.

Resumen visual: lo que necesitas

Distribución	¿Qué le das al Universo?	Nivel
Uniforme	Límites $[a, b]$	Primitivo (ignorancia)
Exponencial	Límite 0 + media	Básico (escala)
Geométrica	Discreto + media	Básico (escala)
Normal	Infinito + media + varianza	Avanzado (energía definida)
Binomial	Probabilidad + regla de parada (n)	Arquitectónico (diseño)

1.4 Distribución binomial negativa

1.4.1 Memoria

Entre evento y evento sí hay falta de memoria: si estás esperando el tercer éxito, el tiempo desde el segundo hasta el tercero es una Geométrica pura (sin memoria). Entonces, ¿por qué la Binomial Negativa (la suma total) **tiene** memoria? Porque tienes una **meta** (*quota*).

Geométrica ($r = 1$): tu meta es conseguir 1. Cada instante es una oportunidad nueva; si fallas, sigues necesitando 1. El pasado no cambia tu necesidad futura inmediata.

Binomial Negativa ($r = 5$): tu meta es conseguir 5. Si llevas 10 minutos y has conseguido 4 éxitos, estás a punto de acabar (te falta 1); si has conseguido 0, estás lejísimos (te faltan 5). La “memoria” aquí no es que el sistema físico (la moneda) cambie, sino que tu **estado de progreso** cambia: la distribución sabe cuánto te falta.

Visualización de la derivada (pendiente)

Sin memoria (Exponencial/Geométrica): la curva siempre cae (e^{-x}). Si estás en $x = 10$, la curva hacia delante es idéntica a la original, solo que más baja. La pendiente relativa $f'(x)/f(x)$ es constante ($-\lambda$): el riesgo de terminar ahora es el mismo que al principio.

Con memoria (Binomial Negativa $r > 1$): la curva tiene una joroba (sube y baja). Al principio (pendiente positiva) es improbable acabar pronto porque necesitas acumular r éxitos; en la joroba está el momento más probable de terminar. Si cortas la gráfica a la mitad, la forma resultante (una cuesta abajo) ya no se parece a la original (una montaña). Eso es tener memoria: el futuro condicionado no se parece al problema original.

Para la propiedad de normalización se usa el binomio generalizado de Newton. Si el exponente es entero (positivo o negativo), existe una demostración combinatoria (contando cosas), la más elegante. Si el exponente es un número raro (fracción, irracional...), no tiene sentido combinatorio inmediato (no puedes “elegir 0.5 bolas”) y la demostración reina es la de cálculo (Taylor).

Para el caso negativo, sabemos que la serie geométrica básica es:

$$(1 - x)^{-1} = \frac{1}{1 - x} = 1 + x + x^2 + x^3 + \dots$$

Esto significa: tengo *una* caja infinita de potencias de x y puedo elegir x^0 , o x^1 , o x^2 ... Ahora, $(1 - x)^{-r}$ significa que tienes r cajas idénticas multiplicándose:

$$(1 + x + x^2 + \dots) \cdot (1 + x + x^2 + \dots) \cdots (r \text{ veces})$$

Para obtener el término x^k en el resultado final, tienes que elegir x^{a_1} de la primera caja, x^{a_2} de la segunda, $\dots$, x^{a_r} de la r -ésima, tal que la suma de exponentes sea k :

$$a_1 + a_2 + \dots + a_r = k, \quad a_i \geq 0 \text{ entero}$$

La pregunta combinatoria. ¿De cuántas formas puedo repartir el número k entre r cajas distintas? Esto es el problema clásico de *stars and bars* (barras y estrellas): tienes k bolas (estrellas) y quieres separarlas en r grupos usando $r - 1$ separadores (barras). El total de objetos a ordenar es k bolas más $(r - 1)$ barras, es decir, $k + r - 1$ posiciones, y tienes que elegir dónde poner las k bolas:

$$\binom{k + r - 1}{k}$$

Conclusión: el coeficiente de x^k en $(1 - x)^{-r}$ es literalmente de cuántas formas puedes sumar k usando r enteros no negativos.

El error clásico: por qué $r^k/k!$ no vale

Es el error que comete todo el mundo (incluidos matemáticos) hasta que se da de bruces con él. La lógica de “elijo caja 1, elijo caja 2 y divido por $k!$ ” funciona **solo** si las cajas elegidas son **distintas**. El problema es cuando repites caja (cuando hay colisiones): ahí la fórmula se rompe y “mata” opciones que existen.

El contraejemplo mortal ($r = 2$ cajas, $k = 2$ objetos). Imagina cajas A y B y 2 bolas idénticas. **La realidad (lo que queremos contar):** ¿de cuántas formas pueden caer? Las dos en A (AA), las dos en B (BB), o una en cada una (AB). Total real: 3 formas.

El método que falla: bola 1 elige caja (2 opciones), bola 2 elige caja (2 opciones), total ordenado $2 \times 2 = 4$: (A, A), (B, B), (A, B), (B, A). Paso final: “como el orden no importa, dividido por $k! = 2$ ”, y sale $4/2 = 2$. ¡Error! La realidad es 3.

La autopsia del error. Al dividir por 2 asumes que **todas** las configuraciones aparecen duplicadas en la lista ordenada (como (A, B) y (B, A)). El par (A, B)/(B, A) es el mismo caso mezclado: hay 2, y al dividir por 2 queda 1 (correcto). Pero el caso (A, A) solo aparece una vez: no tiene “hermano gemelo” al revés. Al dividir su contribución por 2, estás contando medio escenario, y no puedes contar medio escenario. Tu fórmula penaliza injustamente a los “clones” (cuando repites caja).

De dónde sale el $(r + 1)$. Para corregir la fórmula, separamos los casos en los que **no** repites caja de los que **sí**. Seguimos con r cajas y $k = 2$ bolas.

Caso 1 (cajas distintas): bola 1 tiene r opciones, bola 2 tiene $r - 1$ (distinta). Como son distintas, el orden (A, B)/(B, A) molesta y **sí** dividimos por 2. Total: $\frac{r(r-1)}{2}$.

Caso 2 (misma caja, los clones): bola 1 tiene r opciones, bola 2 tiene 1 (la misma). Como es la misma caja, el orden no existe y no hay nada que dividir. Total: r .

Sumando ambos:

$$\text{Total} = \frac{r(r-1)}{2} + r = \frac{r^2 + r}{2} = \frac{r(r+1)}{2}$$

Ahí tienes el término extra: tu fórmula $\frac{r^2}{2}$ venía de $\frac{r \cdot r}{2}$, mientras que la correcta $\frac{r(r+1)}{2}$ añade la corrección por haber “machacado” los casos donde repites caja.

Conclusión final. La intuición de $r^k/k!$ solo vale para muestreo **sin reposición** (o cuando r es casi infinito y la probabilidad de repetir caja es nula). En la Binomial Negativa, repetir caja es crucial: puedes fallar muchas veces antes de acumular los r éxitos. Por eso el método de los separadores (barras y estrellas) es tan genial: gestiona automáticamente las repeticiones sin sumar casos “con” y “sin” repetición a mano. Convierte el problema de meter bolas en cajas en el de ordenar símbolos en una fila, donde las repeticiones no rompen la matemática.

Imagina que quieres conseguir $r = 3$ éxitos (3 caras). El tiempo se divide “físicamente” en 3 etapas o bloques:

- **Bloque 1:** todo lo que pasa hasta que sale la 1.^a cara.
- **Bloque 2:** todo lo que pasa desde la 1.^a cara hasta la 2.^a.
- **Bloque 3:** todo lo que pasa desde la 2.^a cara hasta la 3.^a (fin del juego).

Concepto real	Concepto combinatorio (cajas)	Concepto algebraico (potencias)
Los r éxitos	Son las r cajas	Son los r paréntesis que se multiplican
Los fallos (cruces)	Son las k bolas	Son el exponente de la x (x^k)
Fallos en el bloque 1	Bolas metidas en la caja 1	Exponente elegido en el 1. ^{er} paréntesis (x^{a_1})

Por qué repetir caja importa (interpretación física). Si *no* pudieras repetir caja (si cada caja solo admitiera 1 bola), antes de cada éxito solo podrías tener 0 o 1 fallo: sería un experimento rarísimo donde está prohibido fallar dos veces seguidas. Como *sí* puedes repetir caja (puedes meter 50 bolas en la caja 1), puedes tener 50 fallos seguidos antes de tu primer éxito. Eso es la Binomial Negativa real: puedes tener rachas de mala suerte infinitas.

Conclusión: el coeficiente $\binom{k+r-1}{k}$ cuenta de cuántas formas distintas puedes distribuir tus k fallos a lo largo de los r tiempos de espera: ¿todos los fallos al principio (todas las bolas en la caja 1)?, ¿un fallo entre cada éxito (una bola en cada caja)?, ¿todos los fallos al final (todas las bolas en la caja r)?

1.5 Distribución hipergeométrica

SÍ, ABSOLUTAMENTE. La distribución Hipergeométrica es una probabilidad condicionada. De hecho, esa es la forma más “adulta” y profesional de derivarla.

Olvida las urnas un momento y prepárate para ver la estructura oculta.

1. El experimento mental: los dos cuartos separados

Para ver por qué tu intuición de “producto de independientes” era correcta (pero le faltaba una pieza), imagina este escenario:

Tienes dos habitaciones separadas e insonorizadas.

- **Habitación 1:** contiene N_1 bolas rojas.
- **Habitación 2:** contiene N_2 bolas blancas.

Entras en la Habitación 1 y decides, para cada bola, si te la llevas o no lanzando una moneda (probabilidad π). Esto es una Binomial (N_1, π) . Digamos que te llevas X bolas rojas.

Entras en la Habitación 2 y haces lo mismo. Esto es una Binomial (N_2, π) . Digamos que te llevas Y bolas blancas.

¿Son X e Y independientes? Sí: lo que pase en la Habitación 1 no afecta a la 2. La probabilidad conjunta es el producto:

$$P(X = k \cap Y = m) = P(X = k) P(Y = m).$$

2. El "Dios" del filtro (la condición)

Ahora sales al pasillo con tus bolas rojas y blancas. Llega el "Dios de la Estadística" (o el jefe del experimento) y te dice:

"Solo me interesan los casos en los que te has llevado exactamente n bolas en total ($X + Y = n$). Si te has llevado $n - 1$ o $n + 1$, descarto ese universo paralelo".

Aquí nace la probabilidad condicionada.

Estamos preguntando: ¿cuál es la probabilidad de tener k bolas rojas, dado que la suma total de bolas es n ?

$$P(X = k \mid X + Y = n) = \frac{P(X = k \cap X + Y = n)}{P(X + Y = n)}.$$

Como $X + Y = n$ implica $Y = n - k$, la intersección se reescribe como:

$$P(X = k \mid X + Y = n) = \frac{P(X = k \cap Y = n - k)}{P(X + Y = n)}.$$

Como son independientes (cuartos separados), el numerador se separa:

$$P(X = k \mid X + Y = n) = \frac{P(X = k) P(Y = n - k)}{P(X + Y = n)}.$$

3. La magia algebraica (donde todo encaja)

Si desarrollas esto usando la fórmula de la Binomial ($\binom{N}{x} \pi^x (1 - \pi)^{N-x}$), ocurre un milagro: los términos π y $(1 - \pi)$ se cancelan porque aparecen igual arriba y abajo (la suma de exponentes es n en ambos casos).

Lo que queda es:

$$\frac{\binom{N_1}{k} \binom{N_2}{n-k}}{\binom{N_1+N_2}{n}}.$$

¡BOOM! La Hipergeométrica.

4. La interpretación "bestial"

Tu intuición del "producto" era correcta porque el numerador representa dos procesos independientes que ocurrieron en el vacío, antes de ser observados.

- $\binom{N_1}{k}$ es la "potencia" del evento en la Habitación 1.
- $\binom{N_2}{n-k}$ es la "potencia" del evento en la Habitación 2.

Se multiplican porque, en el espacio de fases original, eran independientes.

Pero la división por $\binom{N}{n}$ es la condición: el colapso que descarta todas las realidades donde no sacaste n bolas en total.

Conclusión La Hipergeométrica es lo que queda de dos Binomiales independientes cuando las fuerzas a cumplir una ley de conservación ($X + Y = n$).

Esto es exactamente la estructura típica de física estadística: partículas moviéndose independientemente, pero impones una conservación (energía/masa), y la distribución pasa de independiente a condicionada.

Intuición física: concentración, moda y truncamiento

Imagina que el conjunto total N es un fluido con una concentración de tinte rojo $p = N_1 / N$.

Cuando sacas una cuchara (muestra n), la naturaleza intenta que la cuchara tenga el mismo sabor (misma concentración) que la olla grande.

La moda Está en

$$k \approx n \cdot \frac{N_1}{N},$$

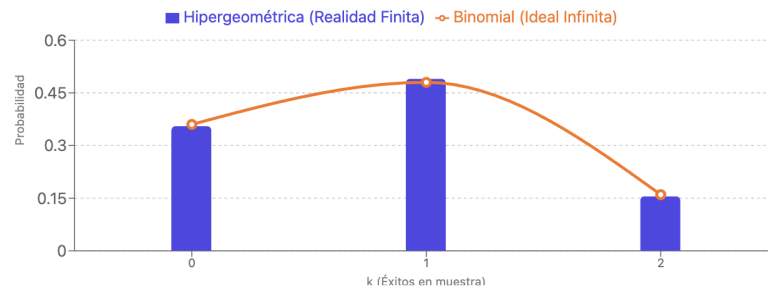
es decir, $k \approx n p$.

La proporción en la muestra (k/n) tiende a imitar la proporción en la población (N_1 / N).

La gráfica Es una campana, sí, pero es una campana truncada y discreta: no puede ir más allá de N_1 (no puedes sacar más éxitos de los que existen) ni menos de $n - N_2$ (si sacas muchas bolas y se acaban los fracasos, a la fuerza tienes que empezar a sacar éxitos).

Visualización geométrica Imagina un histograma que se quiere parecer a una Gaussiana, pero está encerrado entre dos paredes de hormigón: los límites físicos de las bolas disponibles.

Compara la realidad finita (Barras Azules) vs. la idealización infinita (Línea Naranja).

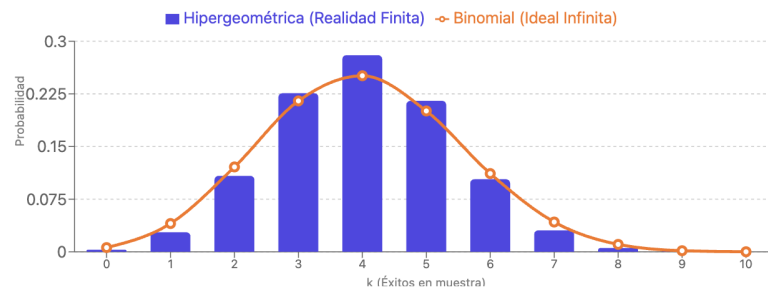


FÍSICA DEL SISTEMA

Media (Esperanza): 0.80
Varianza Hipergeométrica: 0.4702
Varianza Binomial (Referencia): 0.4800
Factor de Corrección Finita: 0.9796
Esto mide cuánta "entropía" has matado al muestrear. Si es 0, el sistema es determinista.

Interpretación en Vivo
El tamaño de muestra es muy pequeño relativo al universo (4.0%). La Hipergeométrica (Azul) y la Binomial (Naranja) son casi idénticas. Apenas has alterado el sistema al sacar bolas.

Compara la realidad finita (Barras Azules) vs. la idealización infinita (Línea Naranja).



FÍSICA DEL SISTEMA

Media (Esperanza): 4.00
Varianza Hipergeométrica: 1.9592
Varianza Binomial (Referencia): 2.4000
Factor de Corrección Finita: 0.8163
Esto mide cuánta "entropía" has matado al muestrear. Si es 0, el sistema es determinista.

Interpretación en Vivo
Estás en la zona intermedia. Se nota la diferencia entre el modelo con reemplazo (Naranja) y sin reemplazo (Azul). La "fricción" de la materia finita empieza a notarse: la línea naranja es más ancha y baja.



Lo que dices es absolutamente correcto: la Hipergeométrica es una Binomial cuando llevas el universo (N) al infinito, manteniendo la proporción ($p = N_1 / N$) constante.

Vamos a explicar esa frase de "la distribución de la información finita" con una analogía de resolución de imagen y luego daremos la demostración matemática paso a paso.

1. La intuición: el universo pixelado vs. vectorial Imagina una fotografía.

La Hipergeométrica (universo pequeño N): es una imagen de baja resolución (pixelada). Si haces zoom y quitas un píxel rojo, la imagen cambia visiblemente. El "porcentaje de rojo" de la foto ha cambiado drásticamente porque había pocos píxeles. Aquí la información es finita y frágil: cada extracción destruye una parte significativa de la realidad.

La Binomial (universo infinito $N \rightarrow \infty$): es una imagen vectorial o de resolución infinita. Si quitas un punto rojo, no pasa nada: hay infinitos puntos rojos. La "densidad" de rojo (p) se mantiene intacta.

Información finita En la Hipergeométrica, cada dato que obtienes (cada bola que sacas) consume la información del sistema: el sistema se "gasta". En la Binomial, el sistema es inagotable.

El objetivo Demostrar que $P(X = k)_{\text{Hiper}} \approx P(X = k)_{\text{Binomial}}$ cuando $N \rightarrow \infty$.

Definimos:

- N : población total ($\rightarrow \infty$).
- K : éxitos en población (N_1). Hacemos que $K \rightarrow \infty$ tal que $K/N = p$ (constante).
- n : muestra (fija y pequeña comparada con N).
- k : éxitos en muestra (fijo).

La fórmula hipergeométrica es:

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}.$$

Desglosemos los combinatorios. Recuerda que

$$\binom{A}{b} = \frac{A(A-1)\dots(A-b+1)}{b!}.$$

Si A es gigante y b es pequeño, podemos decir que $A(A-1)\dots \approx A^b$.

Paso 1: expandir los combinatorios

$$\binom{K}{k} \approx \frac{K^k}{k!}, \quad \binom{N-K}{n-k} \approx \frac{(N-K)^{n-k}}{(n-k)!}, \quad \binom{N}{n} \approx \frac{N^n}{n!}.$$

Paso 2: sustituir en la fórmula original

$$P(X = k) \approx \frac{\left(\frac{K^k}{k!}\right) \left(\frac{(N-K)^{n-k}}{(n-k)!}\right)}{\frac{N^n}{n!}}.$$

Paso 3: reordenar los términos (aquí ocurre la magia) Agrupemos los factoriales para formar el coeficiente binomial y las potencias de N :

$$= \frac{n!}{\underbrace{k!(n-k)!}_{\binom{n}{k}}} \cdot \frac{K^k (N-K)^{n-k}}{N^n}.$$

Paso 4: romper el N^n Sabemos que $N^n = N^k \cdot N^{n-k}$. Repartimos esto debajo de los numeradores:

$$= \binom{n}{k} \cdot \underbrace{\left(\frac{K}{N}\right)^k}_{p^k} \cdot \underbrace{\left(\frac{N-K}{N}\right)^{n-k}}_{(1-K/N)^{n-k}}.$$

Paso 5: identificar los límites Como $K/N = p$:

$$\frac{K}{N} \rightarrow p, \quad \frac{N-K}{N} = 1 - \frac{K}{N} \rightarrow 1 - p = q.$$

Resultado final

$$\lim_{N \rightarrow \infty} P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

¡Voilà! Es la Binomial.

Conclusión "bestial" Lo que has demostrado es que la Binomial es una mentira útil.

La Binomial es un modelo teórico para cuando somos demasiado vagos para contar exactamente cuántas partículas hay en el universo, o cuando el universo es tan grande que no nos importa.

La Hipergeométrica es la física real (materia finita).

La Binomial es la física idealizada (materia infinita/continuo).

Así que sí: la Hipergeométrica es la madre. La Binomial es solo su sombra cuando la proyectas sobre una pared infinita.

1. La Intuición: El Escáner de Resonancia Magnética (MRI) Imagina que la Probabilidad Conjunta (X, Y) es una Montaña en un mapa 3D.

El suelo es el plano XY .

La altura de la montaña en cada punto es la probabilidad (o densidad) $f(x, y)$.

¿Qué es condicionar a $X = x_0$? Es coger una guillotina gigante (un plano) y cortar la montaña exactamente en la coordenada x_0 .

El Corte (La Intersección): Al cortar, te quedas con una "rebanada" de la montaña. Esa rebanada es una curva 2D. Eso es $f(x_0, y)$ (fíjate que la x ya no es variable, es fija).

El Problema: Esa rebanada es pequeña. El área bajo esa curva no suma 1 (suma la probabilidad total de que $X = x_0$, que es pequeña).

El Estiramiento (La Normalización): Para que sea una probabilidad de verdad, tienes que coger esa rebanada y estirla hacia arriba hasta que el área bajo la curva sea 1.

¿Por cuánto divides para estirar? Divides por el área total de la rebanada original ($f_X(x_0)$ o $P(B)$).

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f_X(x)}$$

$f(x, y)$: La forma de la rebanada.

$f_X(x)$: El factor de escala para que el área sea 1.

Conclusión Visual: La probabilidad condicionada es una Tomografía (un corte transversal) del universo, re-escalada para que encaje en nuestra nueva realidad.

2. ¿En qué distribuciones es esto "Bestial"? Crees que solo aplica a la Hipergeométrica, pero te equivocas. Aplica a todas, y en algunas produce resultados mágicos que los físicos aman.

A. La Hipergeométrica (El Corte Diagonal) Lo veremos después. El espacio es una cuadrícula. La condición "Muestra = n " es un plano diagonal ($x + y = n$) que corta el espacio.

Utilidad: Entender que estás restringido a vivir en esa línea diagonal.

B. La Poisson (El Milagro de la Suma) Aquí es donde la geometría se vuelve magia negra. Imagina dos procesos de Poisson independientes (ej. llamadas que llegan a dos teléfonos, X e Y).

El espacio 3D es una montaña formada por $e^{-\lambda_1} \dots e^{-\lambda_2} \dots$

Si lanzas el corte diagonal "Sé que han llegado n llamadas en total ($X + Y = n$)":

Cortas la montaña de Poisson con un plano diagonal.

¿Qué forma tiene la rebanada resultante?

¡Es una Binomial!

$$X \mid (X + Y = n) \sim \text{Binomial}(n, p)$$

(Donde $p = \lambda_1 / (\lambda_1 + \lambda_2)$).

Intuición: El caos absoluto (Poisson) visto a través de un corte de "suma fija" se convierte en orden combinatorio (Binomial).

C. La Normal Multivariante (El Corte de la Campana) Imagina una Campana de Gauss gigante (un sombrero mexicano) en el suelo.

Si la cortas con un cuchillo recto (condicionas $X = x_0$):

La forma de la rebanada sigue siendo una Campana de Gauss.

Esto es único. La Normal es la única distribución que, si la cortas, siempre te da otra Normal (más pequeña y desplazada).

D. La Exponencial (El Corte de la Cola) La propiedad de "falta de memoria" ($P(X > t + s \mid X > t)$) es puramente geométrica.

La curva exponencial cae suavemente.

Si cortas la gráfica en $X = t$ y tiras lo anterior, y luego estiras la parte restante...

¡Se ve idéntica a la original!

Es un fractal. El corte no cambia la forma, solo la escala.

El Escenario: Las Dos Habitaciones Tienes dos habitaciones separadas. No se ven entre ellas.

Habitación Roja: Hay un montón de piezas rojas. Entrás y, por puro azar (lanza una moneda por cada pieza), decides coger k piezas rojas.

Habitación Azul: Hay un montón de piezas azules. Entrás y, por puro azar, decides coger m piezas azules.

¿Qué pasa aquí? Son dos procesos independientes. Lo que hagas en la roja no afecta a la azul.

Si dibujas un mapa (una cuadrícula) de todas las posibilidades:

- Eje X: Cuántas rojas cogiste.
- Eje Y: Cuántas azules cogiste.

En cada casilla (x, y) pones una torre de Legos tan alta como la probabilidad de que eso ocurra. Tienes una montaña 3D en tu mapa.

El "Corte" (La Hipergeométrica) Ahora sales al pasillo con tus bolsas de piezas. Pero hay un guardia de seguridad en la puerta de salida. Te dice:

"¡ALTO! Solo dejo salir a la gente que lleve EXACTAMENTE 10 piezas en total ($n = 10$)."

Aquí aparece la geometría: Tu tenías piezas Rojas (x) y Azules (y).

La regla del guardia es una ecuación: $x + y = 10$.

En tu mapa 2D, $y = 10 - x$. ¡Eso es una línea recta diagonal!

El guardia coge un rayo láser gigante y quema todo el mapa excepto esa línea diagonal.

Todas las posibilidades donde cogiste 11 piezas (ej. 6 rojas, 5 azules) → DESINTEGRADAS.

Todas las posibilidades donde cogiste 9 piezas → DESINTEGRADAS.

Solo sobrevive la "rebanada" diagonal donde la suma es 10.

¿Qué es la Hipergeométrica? Es mirar esa rebanada diagonal superviviente de lado.

Ya no miras todo el mapa. Solo miras la línea del corte.

¿Cuál es la probabilidad de tener 3 rojas? Pues miras la altura de la torre en la casilla $(3, 7)$ (porque $3 + 7 = 10$).

¿Cuál es la probabilidad de tener 4 rojas? Miras la torre en la casilla $(4, 6)$.

¿Por qué narices es útil ver esto así? Porque explica la fórmula "fea" de la Hipergeométrica que tienes en tu PDF sin tener que memorizarla.

$$P(X = k) = \frac{\binom{N_1}{k} \binom{N_2}{n-k}}{\binom{N}{n}}$$

Mírala con "ojos de corte":

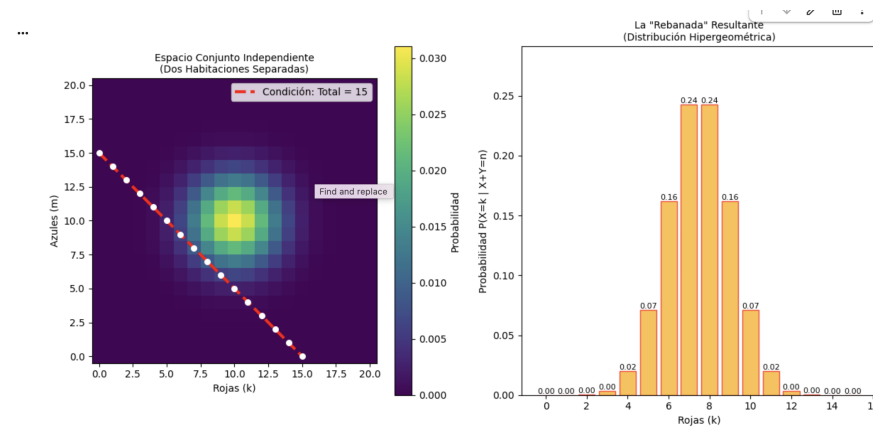
El Numerador ($\binom{N_1}{k} \binom{N_2}{n-k}$): Es el PRODUCTO de dos cosas.

$\binom{N_1}{k}$: "Formas de coger k rojas en la habitación 1".

$\binom{N_2}{n-k}$: "Formas de coger el resto en azules en la habitación 2".

¿Por qué se multiplican? ¡Porque eran independientes antes del corte! En el mapa completo, la altura en (x, y) es $\text{AlturaX} \times \text{AlturaY}$.

El Denominador ($\binom{N}{n}$): Es la suma total de todas las torres que sobrevivieron al láser. Es el factor de normalización para que la "rebanada" sea una probabilidad válida (sume 100).



Si quitas la filosofía y te quedas con la mecánica El denominador no es "la probabilidad de la condición". El denominador es simplemente un número que hace que la suma dé 1.

El Numerador ($f(x, y)$): Contiene la FORMA y la INFORMACIÓN Te dice dónde está la cima de la montaña, dónde baja, si es simétrica, si tiene dos picos...

Es la "rebanada" cruda que sacaste con el láser.

Problema: Su área total es pequeña (ej. 0.03). No es una distribución válida (el área debe ser 1).

El Denominador ($f_X(x)$): Contiene el TAMAÑO o la ESCALA Es un número fijo (una constante para esa rebanada).

Es exactamente el área bajo la curva del numerador: $\int_{-\infty}^{\infty} f(x, y) dy$.

Función: Es el Factor de Corrección. Si tu área es 0.03, divides todo por 0.03.

$\frac{0.03}{0.03} = 1$. ¡Magia! Ya es una probabilidad.

Intuición Bestial: La probabilidad condicionada es como coger una foto pequeña (la rebanada) y hacerle Zoom hasta que ocupe toda la pantalla (probabilidad 1).

El denominador es simplemente el botón de Zoom.

¿Por qué es útil esta visión "Irrespetuosa"? Porque te permite hacer trucos mentales muy potentes, especialmente en Inferencia Bayesiana (que verás más adelante o en otras asignaturas).

La Regla de Oro del "Proporcional A": A menudo, calcular el denominador (esa integral marginal) es una pesadilla de cálculo.

Pero si solo te interesa saber cuál es el valor más probable (la moda), ¡puedes ignorar el denominador!

$$f(y|x) \propto f(x, y)$$

(La condicionada es PROPORCIONAL a la conjunta)

Si $f(x, y)$ tiene un pico en $y = 5$, entonces $f(y|x)$ TAMBIÉN tiene un pico en $y = 5$.

El denominador solo "estira" la gráfica hacia arriba o abajo, no cambia dónde están los máximos ni la forma de la curva. Verlo como un factor de normalización es la forma más

madura, física y geométrica de entenderlo. Olvida la definición de "intersección partido probabilidad total" del colegio. Piensa en "Forma re-escalada".

La demostración combinatoria con bolas de la identidad de Vandermonde es trivial. Podemos rizar el rizo con dos demostraciones interesantes.

1. La Elegancia Algebraica: El Argumento de la Convolución Esta es la demostración que te da una sensación de "poder absoluto". Olvida las personas y los comités. Vamos a usar la estructura misma del Binomio de Newton.

Sabemos que:

$$(1+x)^N = \sum_{k=0}^N \binom{N}{k} x^k.$$

Esto significa que $\binom{N}{k}$ no es solo un número combinatorio: es el coeficiente que acompaña a x^k en ese polinomio. Es su "etiqueta".

La Demostración en 3 líneas

La Identidad Trivial Es obvio que:

$$(1+x)^{N_1} \cdot (1+x)^{N_2} = (1+x)^{N_1+N_2}.$$

Expandimos ambos lados.

Derecha El coeficiente de x^n en $(1+x)^{N_1+N_2}$ es, por definición, $\binom{N_1+N_2}{n}$.

Izquierda Aquí tenemos el producto de dos polinomios:

$$\left(\sum_j \binom{N_1}{j} x^j \right) \cdot \left(\sum_m \binom{N_2}{m} x^m \right).$$

La Pregunta del Francotirador ¿Cómo se forma un término con potencia x^n al multiplicar estos dos paréntesis gigantes?

Necesitas tomar un término x^k del primer paréntesis y multiplicarlo por un término x^{n-k} del segundo paréntesis (porque $x^k \cdot x^{n-k} = x^n$).

Para obtener el coeficiente total de x^n , sumas todas las parejas posibles que sumen n :

$$\text{Coeficiente de } x^n = \sum_{k=0}^n \left[\text{Coef. de } x^k \text{ en el 1º} \right] \times \left[\text{Coef. de } x^{n-k} \text{ en el 2º} \right] = \sum_{k=0}^n \binom{N_1}{k} \binom{N_2}{n-k}.$$

Conclusión Como los polinomios son idénticos, sus coeficientes deben ser idénticos:

$$\binom{N_1+N_2}{n} = \sum_{k=0}^n \binom{N_1}{k} \binom{N_2}{n-k}.$$

¿Por qué es elegante? Porque no cuenta historias. Utiliza la propiedad de convolución de las series de potencias: transforma un problema de conteo en un problema de multiplicación simple.

2. La Elegancia Geométrica: El "Checkpoint" en la Retícula Esta conecta con lo que hablamos antes del "corte", pero visto como un viaje en el tiempo.

Imagina una cuadrícula infinita. Te mueves solo hacia la Derecha (R) o hacia Arriba (U).

$\binom{A+B}{A}$ representa el número de caminos posibles para ir del punto $(0,0)$ al punto (A,B) .

Da igual el camino que tomes, piénsalo un segundo:

Para llegar a tu destino, OBLIGATORIAMENTE vas a tener que dar:

- Exactamente A pasos a la derecha.
- Exactamente B pasos hacia arriba.

Si das menos, no llegas. Si das más, te pasas.

Por lo tanto, la Longitud Total de tu viaje siempre será la suma:

$$\text{Pasos Totales} = A + B.$$

(Es igual que la visión stars and bars)

El Reto Queremos ir del Origen $(0,0)$ al Destino $(n, N_1 + N_2 - n)$.

El número total de pasos es:

$$n + (N_1 + N_2 - n) = N_1 + N_2.$$

El número total de caminos es $\binom{N_1+N_2}{n}$ (este es el lado derecho de Vandermonde).

El Truco del "Checkpoint" Imagina que dibujamos una línea diagonal o "muro" en el mapa exactamente después de haber dado N_1 pasos.

Cualquier camino que vaya desde el origen hasta el destino tiene que cruzar obligatoriamente ese muro temporal en el paso N_1 .

¿Dónde puedes estar tras N_1 pasos? Supongamos que de esos N_1 pasos, has dado k pasos a la derecha (y por tanto $N_1 - k$ hacia arriba). Estás en el punto intermedio

$$M_k = (k, N_1 - k).$$

El camino total se rompe en dos etapas independientes:

Etapas 1 Ir de $(0,0)$ a M_k . Pasos totales: N_1 . Pasos a la derecha: k . Caminos: $\binom{N_1}{k}$.

Etapas 2 Ir de M_k al Destino Final $(n, N_1 + N_2 - n)$. Pasos restantes totales: N_2 .

Pasos a la derecha que faltan: si tenías que dar n en total y ya diste k , te faltan $n - k$.

Caminos: $\binom{N_2}{n-k}$.

La Suma Final Como puedes cruzar el muro en cualquier punto k (desde $k = 0$ hasta $k = n$), sumamos todos los puntos de cruce posibles:

$$\sum_k \underbrace{\binom{N_1}{k}}_{\text{Hasta el muro}} \cdot \underbrace{\binom{N_2}{n-k}}_{\text{Desde el muro al final}} = \binom{N_1 + N_2}{n}.$$

1.6 Memoria

La Topología de la Memoria: ¿Por qué el sistema recuerda?

Tu intuición es que la "combinatoria" (múltiples caminos para llegar a un punto) genera memoria. Vamos a formalizar esto físicamente.

1. Geométrica: El Estado Único (Sin Memoria) Imagina un videojuego donde, si mueres, vuelves al Nivel 1 sin armas y sin experiencia.

- **Variable:** número de intentos hasta pasar el nivel.
- **Estado:** siempre estás en "Nivel 1".
- **Memoria:** ninguna. Si llevas 50 intentos fallidos, tu probabilidad de pasarlo ahora es la misma que la del novato que acaba de empezar.
- **Combinatoria:** no hay "caminos" que se bifurquen. Es un bucle cerrado.

2. Binomial Negativa: El Acumulador (Memoria de Logro) Imagina un videojuego con "checkpoints" o puntos de guardado: necesitas pasar 3 fases.

- **Variable:** intentos hasta pasar las 3 fases.
- **Estado:** puedes estar en Fase 1, Fase 2 o Fase 3.
- **Memoria:** si llevas jugando un rato, es probable que ya estés en la Fase 2. El tiempo transcurrido te da información sobre tu progreso interno.
- **Tu intuición:** "hay combinatoria". Sí: hay muchas formas de haber llegado a la Fase 2 (ganar-perder-ganar, perder-ganar-ganar, etc.). Esa historia interna es la memoria.

3. Binomial / Uniforme / Hipergeométrica: El Desgaste (Memoria de Muerte) Imagina un videojuego con tiempo límite o balas limitadas.

- **Variable:** éxitos obtenidos.
- **Estado:** te quedan X balas o T segundos.
- **Memoria:** cada intento consume un recurso finito. El sistema se está muriendo.
- **Hipergeométrica:** consumes el "recurso" de las bolas en la urna. La estructura del universo cambia físicamente con cada extracción.

El mapa mental definitivo

Distribución	¿Cambia el universo?	¿Acumulo progreso?	¿Hay horizonte final?	¿Tiene memoria?	Analogía
Geométrica	NO	NO	NO	NO	Sísifo (bucle)
Exponencial	NO	NO	NO	NO	Átomo radiactivo (eterna juventud)
Binomial Neg.	NO	SÍ	NO	SÍ	Construir un muro ladrillo a ladrillo
Binomial	SÍ (se gasta n)	SÍ	SÍ	SÍ	Examen con tiempo límite
Hipergeométrica	SÍ (se gasta N)	SÍ	SÍ	SÍ	Comerse una caja de bombones
Poisson	NO	NO	NO	NO*	Lluvia cayendo (ruido puro)

Nota sobre Poisson El proceso (tiempos entre gotas) no tiene memoria. La variable "cuántas gotas caen en 1 hora" sí depende del tamaño de la hora (horizonte), pero el mecanismo generador es amnésico.

Conclusión a tu pregunta sobre la "Única Opción"

La Geométrica no tiene memoria no solo porque haya "una opción", sino porque el fracaso es elástico: el fracaso no deja cicatriz.

En la Binomial Negativa, el éxito deja marca (avanzas).

1.7 Distribución Gamma

Paso 1: El Ladrillo Base (La Exponencial) Primero, visualiza una sola exponencial $X_1 \sim \text{Exp}(\lambda)$. Su densidad es:

$$f_1(t) = \lambda e^{-\lambda t}.$$

- λ (**El Disparo**): Es la probabilidad instantánea de que el evento ocurra ahora mismo.
- $e^{-\lambda t}$ (**La Supervivencia**): Es la probabilidad de que el evento no haya ocurrido durante todo el tiempo t anterior. Es el precio que pagas por haber llegado tan lejos sin que pase nada.

Intuición: Es como un francotirador ciego disparando al azar. λ es la frecuencia de disparo, $e^{-\lambda t}$ es tu suerte de seguir vivo hasta el instante t .

Paso 2: La Suma de Dos (El Relevó) Ahora buscamos la densidad de $S_2 = X_1 + X_2$. Queremos que la suma total sea exactamente t . Para que esto ocurra, el primer evento debe suceder en algún momento x (donde $0 < x < t$) y el segundo debe tardar exactamente el resto del tiempo, $(t - x)$.

Matemáticamente, esto es una convolución, pero veámoslo físicamente:

$$f_2(t) = \int_0^t f_1(x) \cdot f_1(t-x) dx$$

Sustituimos las densidades exponenciales:

$$f_2(t) = \int_0^t (\lambda e^{-\lambda x}) \cdot (\lambda e^{-\lambda(t-x)}) dx$$

¡Aquí viene la magia geométrica! Fíjate en los exponentes:

$$e^{-\lambda x} \cdot e^{-\lambda(t-x)} = e^{-\lambda(x+t-x)} = e^{-\lambda t}.$$

Intuición Brutal: La x desaparece del exponente. Esto significa que no importa cuándo ocurra el relevó. Da igual si el primer corredor corrió mucho y el segundo poco, o al revés. La probabilidad de "sobrevivir" el tiempo total t es constante ($e^{-\lambda t}$), sin importar cómo dividas ese tiempo internamente.

Sacamos lo constante fuera de la integral:

$$f_2(t) = \lambda^2 e^{-\lambda t} \int_0^t 1 dx$$

La integral de 1 entre 0 y t es simplemente la longitud del intervalo: t .

$$f_2(t) = \lambda^2 t e^{-\lambda t}$$

¿Por qué aparece la t ? Porque cuanto más largo es el tiempo total t , más oportunidades (más puntos x) hubo para hacer el cambio de relevó. El "espacio de posibilidades" crece linealmente con el tiempo.

Paso 3: La Suma de Tres (El Volumen) Vamos a por $S_3 = S_2 + X_3$. Sumamos lo que ya teníamos (S_2) con una nueva exponencial.

$$f_3(t) = \int_0^t f_2(x) \cdot f_1(t-x) dx$$

Usamos la fórmula que acabamos de hallar para $f_2(x)$:

$$f_3(t) = \int_0^t (\lambda^2 x e^{-\lambda x}) \cdot (\lambda e^{-\lambda(t-x)}) dx$$

De nuevo, la magia de los exponentes:

$$e^{-\lambda x} \cdot e^{-\lambda(t-x)} = e^{-\lambda t}.$$

¡Se cancelan las x otra vez! Sacamos las constantes (λ^3 y $e^{-\lambda t}$):

$$f_3(t) = \lambda^3 e^{-\lambda t} \int_0^t x dx$$

La integral de x es $x^2/2$. Evaluado de 0 a t :

$$f_3(t) = \lambda^3 e^{-\lambda t} \left[\frac{t^2}{2} \right]$$

Intuición Brutal (Geometría):

Con 2 eventos, buscábamos un punto de corte en una línea (longitud t).

Con 3 eventos, buscamos dos puntos de corte en el tiempo para dividirlo en 3 trozos. Esto es equivalente a calcular el área de un triángulo (o un simplex) en el espacio de fases temporal. Por eso el espacio de posibilidades crece como t^2 (área), no como t (longitud).

¿El factor $1/2$? Es porque el orden de los cortes internos no importa, o geométricamente, porque el área de un triángulo bajo la recta $y = x$ es $\text{base} \times \text{altura} / 2$.

Paso 4: La Generalización (El Patrón Gamma) Si sigues integrando para n eventos, cada vez multiplicas por una λ más y integras una potencia de t mayor:

$$n = 1 \rightarrow t^0 / 0!$$

$$n = 2 \rightarrow t^1 / 1!$$

$$n = 3 \rightarrow t^2 / 2!$$

$$n = 4 \rightarrow \int t^2 / 2 \dots = t^3 / (3 \times 2) = t^3 / 3!$$

El patrón para la suma de n exponenciales es:

$$f_n(t) = \frac{(n-1)!}{\lambda^n} t^{n-1} e^{-\lambda t}$$

Esto coincide exactamente con la definición formal que tienes en tus documentos:

$\Gamma(n) = (n-1)!$ para enteros.

La densidad Gamma es $f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x}$.

Resumen de la Intuición Brutal de la Fórmula Mira la fórmula final de la Gamma y lee sus partes como una historia:

$$f(t) = \underbrace{\frac{(n-1)!}{\lambda^n}}_{\text{Normalización}} \cdot \underbrace{t^{n-1}}_{\text{Espacio de Fases}} \cdot \underbrace{e^{-\lambda t}}_{\text{Supervivencia Global}}$$

- $e^{-\lambda t}$: Es la condición dura. Para que el evento n ocurra exactamente en t , el proceso debe haber estado "corriendo" (sobreviviendo) durante todo el tiempo t .
- t^{n-1} : Es el volumen de configuraciones. Necesitas haber tenido $n - 1$ éxitos previos dentro de ese tiempo t . Cuanto más tiempo t tienes, exponencialmente más formas hay de distribuir esos $n - 1$ puntos intermedios.
- $(n - 1)!$: Es el factor de corrección por el orden. Como los eventos son indistinguibles (son todos exponenciales idénticas), dividimos por las permutaciones posibles de los tiempos intermedios para no contar de más.

El Escenario Imagina que el tiempo t es una cuerda de longitud t metros.

Tú estás al final de la cuerda (en el metro t) porque el n -ésimo evento acaba de ocurrir.

La pregunta es: ¿Dónde cayeron los $n - 1$ eventos anteriores?

Ese término t^{n-1} mide "cuánto sitio" tenían esos eventos anteriores para colocarse.

Caso 1: $n = 2$ (1 evento previo) → La Línea **Situación:** Ha ocurrido el evento 2 en el instante t . El evento 1 tuvo que ocurrir en algún momento antes.

Visualización: Tienes una cuerda de longitud t . Tienes que hacer un corte con la tijera en cualquier punto de esa cuerda para marcar dónde ocurrió el evento 1.

¿Cuánto espacio tienes? Tienes toda la longitud de la cuerda.

Matemática: El "espacio" es la Longitud $= t$.

Potencia: t^1 (que es t^{2-1}).

Intuición: Si duplicas el tiempo ($2t$), tienes el doble de sitio para cortar. El espacio crece linealmente.

Caso 2: $n = 3$ (2 eventos previos) → El Triángulo (Área) **Situación:** Ha ocurrido el evento 3 en el instante t . Los eventos 1 y 2 ocurrieron antes (y en orden: primero el 1, luego el 2).

Visualización: Tienes la misma cuerda de longitud t . Ahora tienes que marcar dos puntos en la cuerda.

El Salto Mental: Para visualizar "todas las combinaciones posibles" de dos puntos, los matemáticos usamos un truco: convertimos los dos tiempos (t_1 y t_2) en coordenadas de un mapa 2D.

Imagina un cuadrado de lado t . Cualquier punto dentro del cuadrado representa dos tiempos al azar.

PERO, el evento 1 debe ser antes que el 2 ($t_1 < t_2$). Eso significa que solo nos vale la mitad del cuadrado (la mitad inferior bajo la diagonal).

Esa mitad es un triángulo.

¿Cuánto espacio tienes? Es el Área de ese triángulo.

Matemática: El área de un triángulo con base t y altura t es proporcional a $t \times t = t^2$.

Potencia: t^2 (que es t^{3-1}).

Intuición: Si duplicas el tiempo ($2t$), el área del triángulo se multiplica por 4 (2^2). ¡El espacio de posibilidades crece mucho más rápido que el tiempo!

Caso 3: $n = 4$ (3 eventos previos) $\rightarrow$ **La Pirámide (Volumen)** **Situación:** Ha ocurrido el evento 4 en t . Hubo 3 eventos antes.

Visualización: Tienes que marcar 3 puntos en la cuerda.

El Salto Mental: Si conviertes 3 puntos en coordenadas (x, y, z) , estás en un espacio 3D.

Todas las combinaciones posibles forman un cubo de lado t .

Pero como tienen que ir en orden ($t_1 < t_2 < t_3$), solo nos vale una esquina del cubo. Esa esquina es una pirámide (tetraedro).

¿Cuánto espacio tienes? Es el Volumen de esa pirámide.

Matemática: El volumen escala con "lado x lado x lado", es decir, t^3 .

Potencia: t^3 (que es t^{4-1}).

Intuición: Si duplicas el tiempo ($2t$), el volumen de posibilidades se multiplica por 8 (2^3).

¿Qué es un "Simplex"? Un Simplex es simplemente el nombre pijo para esta forma geométrica que generaliza el triángulo y la pirámide a más dimensiones.

En 1D: Es un segmento de línea.

En 2D: Es un triángulo.

En 3D: Es un tetraedro (pirámide de base triangular).

En 4D: Es un "pentácoron" (hiper-pirámide).

La intuición brutal:

Un simplex es lo que obtienes cuando cortas un "cubo" (o hipercubo) por sus diagonales para quedarte solo con la parte donde las coordenadas están ordenadas ($x < y < z < \dots$).

Resumen: ¿Por qué t^{n-1} ? Cuando esperas al evento n :

Fijas el final en t .

Quedan $n - 1$ eventos "flotando" libremente en el pasado.

Esos $n - 1$ eventos definen un objeto geométrico de $n - 1$ dimensiones (una línea, un triángulo, una pirámide...).

El "tamaño" o "volumen" de ese objeto siempre se calcula elevando el lado (t) a la dimensión del objeto ($n - 1$).

1. El t^{n-1} : "El Abanico de Historias" (La Entropía) Imagina que estás en el instante t . Acaba de ocurrir el evento final, el número n . Miras hacia atrás, hacia el pasado (desde 0 hasta t).

La Pregunta: "¿Cómo diablos hemos llegado hasta aquí?"

La Respuesta: Hubo $n - 1$ eventos anteriores.

El Simplex: Como los eventos son indistinguibles (son puntos anónimos en el tiempo), no nos importa cuál fue el primero, sino simplemente que hubo un primero, un segundo, etc.

El término t^{n-1} no es un volumen de espacio físico. Es un volumen de HISTORIA.

Si el tiempo t es muy corto (digamos, 1 segundo), el "volumen" es minúsculo. Apenas hay sitio para meter la historia de $n - 1$ eventos. Todo está apretado. Hay pocas "versiones" de esa historia.

Si el tiempo t es enorme (un siglo), el "volumen" es gigantesco. Los eventos pudieron ocurrir en el año 1, o en el 50, o todos ayer. El abanico de cronologías posibles se abre descomunadamente.

Intuición Brutal: El t^{n-1} cuenta los Microestados Temporales. Es la medida del desorden. Cuanto más tiempo pasa, más "caótica" y rica puede ser la historia interna de cómo llegaste al final.

2. El e^{-t} : "El Impuesto a la Existencia" (La Supervivencia) Ahora, mira el otro término.

La Realidad: El universo es hostil. El proceso de Poisson es un proceso de "disparo constante".

El Significado: Para que tu historia llegue hasta el tiempo t , el sistema tuvo que no romperse (o no completarse) durante cada milisegundo anterior.

El Decaimiento: e^{-t} es el precio que pagas por durar. Es la "Fricción de la Realidad".

Intuición Brutal: Es como correr bajo una lluvia de flechas. Cuanto más tiempo corres (t), más improbable es que sigas vivo. e^{-t} es la probabilidad de que tu historia no haya sido cortada antes de tiempo.

3. La Fusión: La "Densidad" como un Balance de Poder Aquí es donde la frase cobra vida. La función de densidad $f(t)$ es el producto de esas dos fuerzas titánicas chocando.

$$f(t) \propto \underbrace{\text{Volumen de Historias}(t^{n-1})}_{\text{Posibilidad}} \times \underbrace{\text{Probabilidad de Sobrevivir}(e^{-t})}_{\text{Realidad}}$$

Visualízalo como un Cono de Luz expandiéndose y desvaneciéndose a la vez:

Al inicio: El cono se abre. El volumen (t^{n-1}) explota hacia afuera. Hay muchísimas nuevas formas de acomodar los eventos. La "Posibilidad" gana. La curva sube.

El conflicto: El "Impuesto a la Existencia" (e^{-t}) empieza a cobrar su deuda. La lluvia de flechas es demasiado densa.

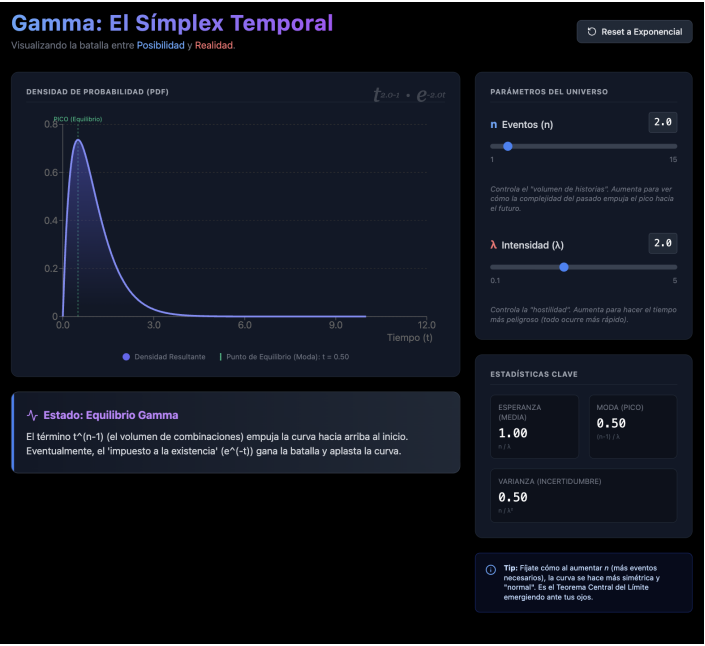
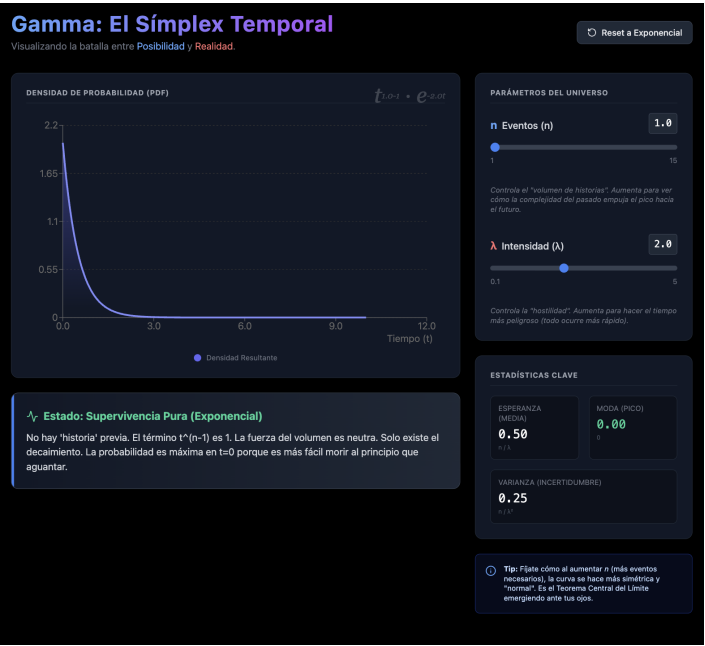
El final: Aunque el volumen de historias posibles es teóricamente infinito si vivieras eternamente, la probabilidad de llegar vivo a la eternidad es cero. La "Realidad" aplasta a la "Posibilidad". La curva cae.

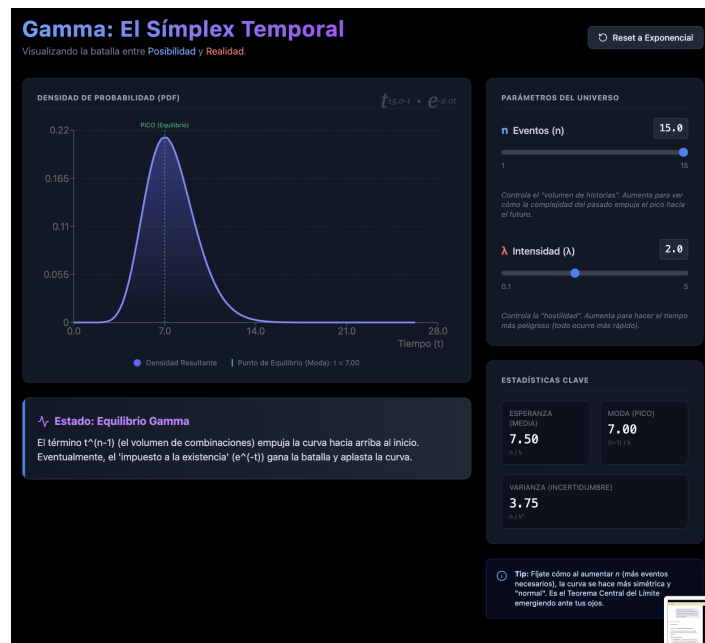
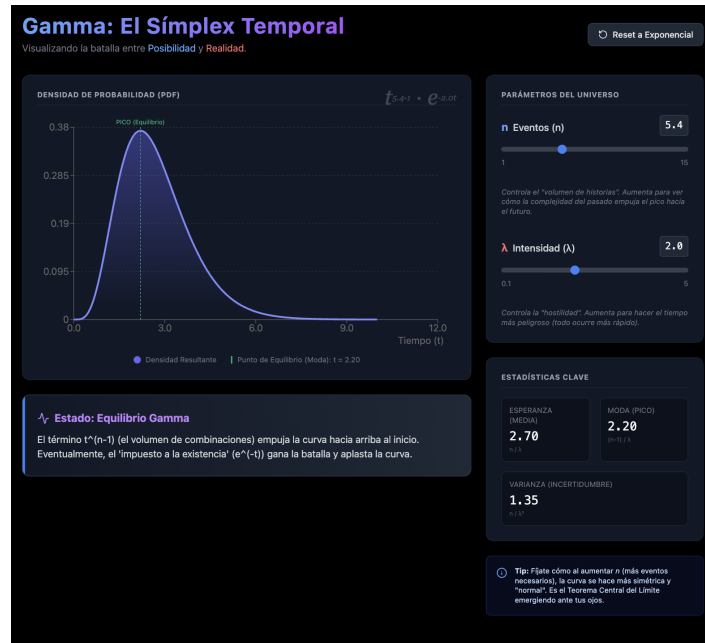
Conclusión Filosófica La función Gamma nos dice que el momento más probable para que termine el proceso (la moda) es exactamente el punto donde:

La ganancia en libertad combinatoria (ganar más espacio para ordenar tu pasado) se equilibra exactamente con el riesgo de extinción (la probabilidad de morir en el futuro).

El "Volumen del Simplex Temporal" es la riqueza de tu pasado, y el "Decaimiento" es la fragilidad de tu futuro. La Gamma es la foto de ese equilibrio.

¡Y así, la extraña función Gamma $t^{n-1}e^{-t}$ se revela simplemente como el volumen de un "simplex temporal" ponderado por la probabilidad de decaimiento!





1. **¿Qué es la Convolución? (La "Mezcladora de Historias")** Formalmente, dada una variable X (con función f) y una variable Y (con función g), la distribución de su suma $S = X + Y$ es la convolución ($f * g$).

Caso Continuo (Integral):

$$h(t) = \int_{-\infty}^{\infty} f(x) \cdot g(t - x) dx$$

Intuición: Cortas el total t en dos trozos: x y $(t - x)$. Barres todas las posibles ubicaciones del corte (la integral).

Caso Discreto (Sumatoria):

$$h[k] = \sum_j f[j] \cdot g[k-j]$$

Intuición: Tienes k fallos totales. Si en la primera etapa hubo j fallos, en la segunda obligatoriamente hubo $k-j$. Sumas la probabilidad de todas esas parejas posibles.

Tu analogía de "bolas en cajas":

Es perfecta. Si tienes que sumar un total de k (bolas) usando r variables (cajas), la convolución cuenta de cuántas formas puedes repartir esas bolas. En matemáticas discretas, esto genera los coeficientes binomiales.

2. El Desarrollo Paralelo: Continuo vs. Discreto Vamos a ver cómo la Gamma nace de la Exponencial (tiempo fluyendo) y cómo la Binomial Negativa nace de la Geométrica (fallos acumulándose), usando los datos de tus documentos.

A. El Ladrillo Base (El evento unitario) Mundo Continuo (Exponencial):

La variable $X \sim \text{Exp}(\lambda)$ mide el tiempo hasta el primer éxito. Su densidad es "supervivencia pura":

$$f(t) = \lambda e^{-\lambda t}$$

Mundo Discreto (Geométrica):

La variable $X \sim \text{Ge}(p)$ mide el número de fracasos hasta el primer éxito. Su probabilidad es "racha de fallos + éxito final":

$$P(X = k) = q^k p$$

(Donde $q = 1 - p$ es la probabilidad de fracaso).

B. La Suma de Dos (El primer "cruce") Queremos sumar dos variables independientes.

Mundo Continuo (Suma de 2 Exponenciales):

Queremos tiempo total t .

$$f_2(t) = \int_0^t (\lambda e^{-\lambda x})(\lambda e^{-\lambda(t-x)}) dx = \lambda^2 e^{-\lambda t} \underbrace{\int_0^t 1 dx}_{\text{Longitud } t}$$

Resultado: $\lambda^2 t e^{-\lambda t}$.

Intuición: El término t representa la "longitud" donde pudimos hacer el relevo.

Mundo Discreto (Suma de 2 Geométricas):

Queremos un total de k fallos acumulados tras conseguir 2 éxitos consecutivos (uno detrás de otro).

$$P(S_2 = k) = \sum_{j=0}^k P(X_1 = j) \cdot P(X_2 = k-j)$$

Sustituimos la fórmula de la geométrica:

$$P(S_2 = k) = \sum_{j=0}^k (q^j p) \cdot (q^{k-j} p)$$

Fíjate en la magia algebraica (idéntica a la exponencial): $q^j \cdot q^{k-j} = q^k$. ¡La j desaparece del exponente!

$$P(S_2 = k) = p^2 q^k \underbrace{\sum_{j=0}^k 1}_{\text{Cuenta } k+1}$$

Intuición: Ese $(k+1)$ es el equivalente discreto de la t . Si tienes k fallos totales, ¿cuántas formas hay de repartirlos entre el primer y segundo intento?

0 fallos en el 1º, k en el 2º.

1 fallo en el 1º, $k-1$ en el 2º.

...

k fallos en el 1º, 0 en el 2º.

Total: $k+1$ posibilidades.

C. La Generalización (La Meta Final) Aquí es donde tu intuición de "bolas en cajas" brilla. **Mundo Continuo (Gamma - Suma de r Exponenciales):**

La Gamma representa el tiempo hasta r éxitos. La fórmula es:

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$$

El término x^{r-1} es el "volumen" del espacio de posibilidades continuas.

Mundo Discreto (Binomial Negativa - Suma de r Geométricas):

La Binomial Negativa $BN(r, p)$ es el número de fallos hasta obtener r éxitos. Es la suma de r geométricas independientes.

La convolución de r geométricas nos obliga a repartir k fallos (bolas) en r periodos de espera (cajas).

El número de formas de hacer esto es el número combinatorio con repetición:

$$\binom{k+r-1}{k}$$

(Nota: Si pones $r = 2$, esto da $\binom{k+1}{k} = k+1$, ¡que es exactamente lo que calculamos en el paso anterior!)

Por tanto, la función de masa es:

$$P(X = k) = \underbrace{\binom{k+r-1}{k}}_{\text{Espacio de Posibilidades}} \cdot \underbrace{q^k p^r}_{\text{Probabilidad Base}}$$

Visualizando la convolución Pinta la función $f(x)$ en el eje horizontal.

Pinta la función $g(y)$ en el eje vertical. Eje horizontal: Tiempo del evento 1 (x).

Eje vertical: Tiempo del evento 2 (y).

La probabilidad conjunta ($P(X) \cdot P(Y)$) forma una nube de densidad sobre este mapa. Como son exponenciales, la nube es muy densa cerca del origen $(0,0)$ y se desvanece al alejarse. El producto $f(x)g(y)$ crea una montaña 3D (una superficie) sobre el plano xy .

La Pregunta: ¿Cuál es la probabilidad de que $S = X + Y = t$?

En el plano xy , la ecuación $x + y = t$ (o $y = -x + t$) es una línea diagonal que corta el plano a 45 grados.

La Convolución es:

Coger un cuchillo y cortar esa montaña 3D siguiendo esa línea diagonal $x + y = t$. El área de la "loncha" de carne que te queda en el cuchillo es el valor de la convolución en t .

- Si t es pequeño: La diagonal pasa cerca del origen $(0,0)$. Cortas la falda de la montaña. Poca área.
- Si t crece: La diagonal se mueve hacia afuera (arriba a la derecha). Cortas por el centro de la montaña. Mucha área.
- Si t es enorme: La diagonal se va muy lejos. Ya no cortas montaña. Área cero.
- **Fijas t (digamos $t = 5$):** Dibujas una línea diagonal que conecta el 5 del eje x con el 5 del eje y .
- **La Integral:** Caminas sobre esa línea. En cada paso, miras cuánta "niebla" (densidad de probabilidad) hay en ese punto.
- **El Resultado:** Sumas toda la niebla que has pisado. Ese es el valor de la densidad en t .

Visualiza la Gamma aquí La "niebla" de dos exponenciales ($e^{-x} \cdot e^{-y}$) tiene una propiedad mágica: en cualquier punto de esa línea diagonal, la densidad es constante (porque $e^{-(x+y)} = e^{-t}$).

Así que caminar por la línea es sumar el mismo valor todo el rato.

El resultado de la integral depende solo de cuánto tiempo caminas (la longitud de la línea).

- Si t es pequeño, la línea diagonal es corta (cerca de la esquina).
- Si t es grande, la línea diagonal es larga.

¡Por eso la probabilidad sube linealmente con t ! Porque la línea diagonal sobre la que integras se hace más larga a medida que t se aleja del origen.

1. LA DEFINICIÓN FORMAL (El "God Protocol") En matemáticas, una Convolución es una operación binaria entre dos secuencias (o funciones) a y b , denotada como $(a * b)$.

Si tienes dos secuencias de números:

$$a = (a_0, a_1, a_2, \dots), \quad b = (b_0, b_1, b_2, \dots).$$

La convolución es una nueva secuencia c , donde el término n -ésimo es:

$$c_n = (a * b)_n = \sum_{k=0}^n a_k \cdot b_{n-k}.$$

Punto y final. Esa es la definición.

Es la suma de productos cruzados donde los índices suman n ($k + (n - k) = n$).

¿Por qué existe esta operación? (La clave algebraica) Esta operación no es arbitraria. Es exactamente lo que ocurre con los coeficientes cuando multiplicas dos polinomios.

Si defines:

$$A(x) = a_0 + a_1x + a_2x^2 + \dots, \quad B(x) = b_0 + b_1x + b_2x^2 + \dots$$

Entonces el producto $C(x) = A(x) \cdot B(x)$ tiene como coeficientes la convolución de a y b .

$$C(x) = \sum_{n=0}^{\infty} \underbrace{(a * b)}_{\text{convolución}}_n x^n.$$

2. APLICACIÓN A TUS DOS CASOS (Desentrañando el lío) Tú ves una relación porque AMBOS casos son convoluciones. La diferencia es QUÉ estás convolucionando: ¿Probabilidades o Formas de Contar?

CASO A: La Binomial Negativa / Gamma (Suma de Variables Aleatorias) Lo que convolucionas: Probabilidades ($P(X = k)$).

El Polinomio: La "Función Generatriz de Probabilidad".

La Binomial Negativa $BN(r, p)$ es la suma de r variables Geométricas independientes.

La función generatriz de una Geométrica es (simplificando): $G(x) = \frac{p}{1-qx}$.

Si sumas variables aleatorias, multiplicas sus polinomios generatrices:

$$G_{\text{total}}(x) = G(x) \cdot G(x) \cdots G(x) = \left(\frac{p}{1-qx} \right)^r.$$

Para encontrar la probabilidad de k fallos (el coeficiente de x^k), tienes que expandir ese polinomio resultante. Al hacerlo, los coeficientes se mezclan mediante convolución.

Resultado: La fórmula de la Binomial Negativa.

Intuición: Aquí la convolución representa "La suma de tiempos o fallos acumulados". El índice n es el tiempo total o fallos totales.

CASO B: La Identidad de Vandermonde / Hipergeométrica (Combinatoria) Lo que convolucionas: Conteos Combinatorios ($\binom{N}{k}$).

El Polinomio: El Binomio de Newton $(1+x)^N$.

Mira la identidad de Vandermonde:

$$\sum_k \binom{r}{k} \binom{m}{n-k} = \binom{r+m}{n}.$$

Esto es, literalmente, una convolución de dos secuencias:

- Secuencia a : los coeficientes de $(1+x)^r \rightarrow \binom{r}{0}, \binom{r}{1}, \dots$
- Secuencia b : los coeficientes de $(1+x)^m \rightarrow \binom{m}{0}, \binom{m}{1}, \dots$

Si multiplicas los polinomios:

$$(1+x)^r \cdot (1+x)^m = (1+x)^{r+m}.$$

El coeficiente de x^n en el lado derecho es $\binom{r+m}{n}$.

El coeficiente de x^n en el lado izquierdo (el producto) es la convolución de los coeficientes individuales.

Intuición: Aquí la convolución representa "Particionar un conjunto fijo". El índice n es el tamaño de la muestra que extraes.

3. EL MAPA MENTAL DEFINITIVO Aquí está la distinción que te falta para que todo encaje. La convolución es la herramienta matemática (el motor). Lo que cambia es la "gasolina" que le echas.

Característica	Suma de Variables Indep.	Muestreo / Partición
¿Qué sumamos?	El valor de la variable aleatoria: $X + Y = Z$.	El tamaño de la sub-muestra: $n_1 + n_2 =$
Objeto Matemático	Función Generatriz de Momentos / Probabilidad.	Polinomio Combinatorio $(1 + x)^N$.
Operación	$(p_1 * p_2)[z]$	$(\text{Combi}_1 * \text{Combi}_2)[n]$
Significado del índice	Magnitud acumulada (Tiempo total, Energía total).	Cantidad de elementos elegidos.
Tu Confusión	"Repartir bolas (fallos) en el tiempo"	"Elegir bolas de una urna"

Visualización geométrica (para que se te quede grabado) Para ver esto geométricamente y que se te quede grabado para siempre, olvida las funciones 1D. Piensa en una Tabla (Matriz).

Caso Discreto (Dados / Polinomios) Imagina que pones la secuencia a en el eje horizontal y la b en el vertical.

Rellenas la tabla multiplicando $a_i \cdot b_j$.

$b \downarrow / a \rightarrow$	a_0	a_1	a_2
b_0	$a_0 b_0$ (Suma=0)	$a_1 b_0$ (Suma=1)	$a_2 b_0$ (Suma=2)
b_1	$a_0 b_1$ (Suma=1)	$a_1 b_1$ (Suma=2)	$a_2 b_1$ (Suma=3)
b_2	$a_0 b_2$ (Suma=2)	$a_1 b_2$ (Suma=3)	$a_2 b_2$ (Suma=4)

¿Qué es la convolución? Es sumar las DIAGONALES INVERTIDAS (las que van de abajo-izquierda a arriba-derecha).

Para $n = 0$: Solo $a_0 b_0$.

Para $n = 1$: Sumas $a_0 b_1 + a_1 b_0$.

Para $n = 2$: Sumas $a_0 b_2 + a_1 b_1 + a_2 b_0$.

Caso Continuo (Mecánica) Ahora haz que esa tabla tenga resolución infinita. Se convierte en un plano continuo xy .

Eje X: Tiempo τ (cuando aplicas la fuerza).

Eje Y: Tiempo Δt (retardo del sistema).

Tienes una montaña de valores $F(\tau) \cdot g(y)$ en ese plano.

¿Qué es la convolución en t ? Es trazar la línea recta DIAGONAL definida por $x + y = t$ (o sea, $y = t - x$) y calcular el área de la pared que se levanta sobre esa línea.

Conclusión Geométrica La convolución es cortar el espacio de posibilidades (el plano xy) con una cuchilla diagonal ($x + y = t$) y sumar todo lo que pillas en el corte.

Es idéntico para discretas y continuas.

En la Hipergeometría, la extracción deja marca (hueco en la urna).

En la Geometría, el sistema se "cura" instantáneamente tras cada fallo: es la única distribución perfectamente elástica.

PASO 1: La Fórmula de Partida (Binomial Negativa)

La probabilidad de tener k fallos antes del r -ésimo éxito es:

$$P(X = k) = \binom{k+r-1}{k} \cdot p^r \cdot (1-p)^k$$

Vamos a diseccionar esto en tres trozos y aplicar el límite ($\Delta t \rightarrow 0$) a cada uno. Sustituimos $k = t/\Delta t$ y $p = \lambda\Delta t$.

PASO 2: El Corazón del Decaimiento $((1-p)^k)$ Este término es la probabilidad de fallar k veces seguidas. Sustituimos:

$$(1 - \lambda\Delta t)^{t/\Delta t}$$

Esto es un límite clásico de cálculo. Recuerda que $\lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n = e^x$. Aquí $n = 1/\Delta t$.

$$\lim_{\Delta t \rightarrow 0} (1 - \lambda\Delta t)^{1/\Delta t \cdot t} = e^{-\lambda t}$$

Intuición Brutal: La probabilidad de fallar en un millón de micro-intentos $((1-p)^k)$ se convierte en la probabilidad de sobrevivir un tiempo continuo ($e^{-\lambda t}$). El “rosario de fallos” se funde en una “espera continua”.

PASO 3: El Motor del Evento (p^r) Este término es la probabilidad de que ocurran los r éxitos necesarios. Sustituimos:

$$p^r = (\lambda\Delta t)^r = \lambda^r (\Delta t)^r$$

Aquí nos sobra un $(\Delta t)^r$. Guárdalo, lo usaremos para limpiar el desastre combinatorio.

PASO 4: El Espacio de Fases (El Combinatorio) Aquí está la magia sucia.

$$\binom{k+r-1}{k} = \frac{(k+r-1)!}{k!(r-1)!}$$

Como $\Delta t \rightarrow 0$, el número de ticks k tiende a infinito (hay billones de nanosegundos en un segundo). Cuando k es gigantesco y r es pequeño (fijo), podemos aproximar el factorial.

Piensa: $\frac{1005!}{1000!} = 1005 \cdot 1004 \cdot 1003 \cdot 1002 \cdot 1001 \approx 1000^5$. Generalizando:

$$\frac{(k+r-1)!}{k!} \approx k^{r-1}$$

Así que el combinatorio se simplifica a:

$$\approx \frac{k^{r-1}}{(r-1)!}$$

Sustituimos $k = t/\Delta t$:

$$\approx \frac{(t/\Delta t)^{r-1}}{(r-1)!} = \frac{t^{r-1}}{(r-1)! (\Delta t)^{r-1}}$$

Intuición Brutal: El número combinatorio (contar formas de ordenar bolas discretas) se ha convertido en un monomio t^{r-1} (volumen geométrico). Lo discreto se ha vuelto geometría.

PASO 5: El Ensamblaje Final (La Fusión) Juntamos las tres piezas transformadas:

1. **Combinatorio:** $\frac{t^{r-1}}{(r-1)!(\Delta t)^{r-1}}$

2. **Motor:** $\lambda^r (\Delta t)^r$

3. **Decaimiento:** $e^{-\lambda t}$

Multiplicamos todo:

$$P(X = k) \approx \left[\frac{t^{r-1}}{(r-1)!(\Delta t)^{r-1}} \right] \cdot [\lambda^r (\Delta t)^r] \cdot [e^{-\lambda t}]$$

¡Mira las Δt ! Tienes $(\Delta t)^r$ arriba y $(\Delta t)^{r-1}$ abajo. Se cancelan todas menos una (Δt) .

$$P(X = k) \approx \underbrace{\left[\frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t} \right]}_{\text{DENSIDAD GAMMA}} \cdot \Delta t$$

LA CONCLUSIÓN BRUTAL

$$P(\text{ticks} = k) \approx f_{\text{Gamma}}(t) \cdot dt$$

Lo que acabamos de ver es alquimia matemática:

- **La Combinatoria** ($\binom{n}{k}$) **se convirtió en Geometría** (t^{r-1}).

De contar formas de ordenar puntos a medir el volumen de un simplex.

- **La Potencia** (p^r) **se convirtió en Escala** (λ^r).

De probabilidad por evento a densidad por unidad de tiempo.

- **El Producto Discreto** ($(1-p)^k$) **se convirtió en Exponencial** ($e^{-\lambda t}$).

De “no fallar tick a tick” a “sobrevivir en el continuo”.

- **El Residuo** (Δt):

El Δt que sobra al final es vital. Convierte la “Probabilidad” (adimensional, suma 1) en “Densidad” (dimensional, integra 1). Nos dice que la probabilidad de un punto exacto en el continuo es 0, y que necesitamos un diferencial dt para tener algo real.

Hemos demostrado que la Gamma no es más que una Binomial Negativa vista a través de unas gafas miopes que no distinguen los ticks individuales del reloj, viendo solo el flujo del tiempo.

La Función Gamma no es solo una integral fea; es el puente místico entre lo discreto (el conteo, los números naturales) y lo continuo (el flujo, el cálculo).

Aquí tienes la disección visceral de la $\Gamma(z)$.

1. El Origen: “Rellenando los huecos de Dios” (1729)

El Problema: Imagina los números factoriales como puntos flotando en el vacío:

$$1! = 1$$

$$2! = 2$$

$$3! = 6$$

$$4! = 24$$

A principios del s. XVIII, Daniel Bernoulli y Christian Goldbach se preguntaron: ¿Qué hay en medio? ¿Cuánto vale $2.5!$? ¿Existe una curva suave que pase por todos esos puntos?

Parece una pregunta inocente, pero es profunda. El factorial es una operación fundamental de la combinatoria (formas de ordenar cosas). Preguntar por el factorial de un número no entero es preguntar: “¿De cuántas formas puedo ordenar 2.5 objetos?”. Parece absurdo.

El Héroe: Leonhard Euler. Euler, el Mozart de las matemáticas, encontró la solución. No buscó una fórmula algebraica simple (no existe). Buscó un proceso dinámico que generara esos números.

Descubrió que la integral (el área bajo una curva que decae) generaba exactamente la secuencia factorial si ajustabas la potencia.

2. La Definición Brutal: La Máquina de Procesar Escalas La definición es:

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

Vamos a leer esta fórmula con visión de rayos X. No es un área, es una batalla entre dos titanes:

t^{z-1} (El Crecimiento Polinómico): Esta parte quiere explotar hacia el infinito. Si z es grande, esta curva sube y sube. Representa la expansión, la entropía, el volumen, las posibilidades creciendo con el tiempo.

e^{-t} (El Decaimiento Exponencial): Esta parte es la guadaña. Mata todo lo que toca a medida que t avanza. Representa la probabilidad, la restricción, la finitud de la vida.

¿Qué es la Gamma intuitivamente? Es la suma total de la energía de un sistema que intenta crecer (t^{z-1}) pero está condenado a morir (e^{-t}).

Si z es pequeño (poca ambición de crecimiento), la exponencial lo mata rápido. El área es pequeña.

Si z es grande (mucha potencia t^{100}), la función aguanta creciendo mucho tiempo antes de que la exponencial la doblegue. El área es gigantesca (100!).

Intuición Física: La Función Gamma mide la “persistencia total” de un proceso de potencia antes de ser aniquilado por la probabilidad.

3. ¿Por qué $\Gamma(n) = (n-1)!$? (El Misterio del -1) Esta es la parte que confunde a todo el mundo. ¿Por qué el factorial de n es $\Gamma(n+1)$? ¿Por qué ese desfase?

La respuesta más “brutal” viene de la Física Estadística y la Geometría Dimensional.

La Intuición Geométrica: El Volumen de la N-Esfera Aquí es donde la Gamma se vuelve la reina del universo.

Si quieres calcular el volumen de una bola en n dimensiones (una hiperesfera):

2D (Círculo): πR^2

3D (Esfera): $\frac{4}{3} \pi R^3$

n D (Hiperesfera): $V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} R^n$

¡La Gamma está en el denominador!

¿Qué hace ahí? La Gamma es el factor que te dice “cuánto espacio se desperdicia en las esquinas” a medida que subes de dimensión.

Un cubo en 100 dimensiones es casi todo “esquinas”. La esfera inscrita en el centro es minúscula.

La función Gamma crece tan rápido ($n!$) que hace que el volumen de la hipersfera tienda a CERO cuando la dimensión tiende a infinito (para radio 1).

Intuición: La Gamma es el regulador dimensional. Mide cómo la geometría se “comprime” o se vuelve más compleja a medida que añades grados de libertad.

Estas son las propiedades que explican por qué la gráfica tiene esa forma y qué significan sus parámetros α (forma) y λ (ritmo).

1. La Propiedad Recursiva: “La Escalera de la Complejidad”

$$\Gamma(x+1) = x \cdot \Gamma(x)$$

¿Qué es matemáticamente? Es la generalización de $(n+1)! = (n+1) \times n!$.

Intuición Brutal para la Distribución: Imagina que la variable Gamma modela el tiempo para completar una tarea compleja de x pasos. Esta propiedad te dice que la complejidad del paso siguiente es multiplicativa, no aditiva. El “volumen de posibilidades” escala con el tamaño del sistema.

Por esto, la Media de la distribución Gamma es x/λ . Es lineal. Pero la Varianza es x/λ^2 .

Esta propiedad es la que permite conectar una Gamma de parámetro 4 con una de parámetro 5. Son “familia”. Puedes construir procesos complejos sumando procesos simples.

Volvamos a los relevos. Si cada corredor tarda de media 10 minutos ($\lambda = 0.1$, esperanza $1/\lambda = 10$). Y tienes 4 corredores ($\alpha = 4$).

¿Cuánto tarda la carrera en total?

$$10 + 10 + 10 + 10 = 40 \text{ minutos.}$$

$$E[Total] = E[C_1] + E[C_2] + E[C_3] + E[C_4] = \frac{1}{\lambda} \times \alpha = \frac{\alpha}{\lambda}$$

¿Por qué es tan simple? Por la Linealidad de la Esperanza. Al tiempo físico no le importa la complejidad combinatoria del “espacio de fases”. El tiempo es aditivo. Si pones 4 bombillas en serie y cada una dura 1000 horas, la cadena dura 4000 horas (de uso secuencial).

Conclusión: Mientras que el “drama” interno de las combinaciones se multiplica, el resultado externo promedio simplemente se suma.

3. La Varianza Sub-lineal ($Var = \alpha/\lambda^2$): “La Cancelación de la Suerte” “La Varianza es α/λ^2 . Esta propiedad conecta Gamma(4) con Gamma(5).”

Esta es la parte más profunda. Fíjate que la Desviación Típica (la incertidumbre real, σ) es la raíz cuadrada de la varianza:

$$\sigma = \frac{\sqrt{\alpha}}{\lambda}$$

La Paradoja: La Media crece con α (linealmente, paso a paso 1, 2, 3, 4...). La Incertidumbre (σ) crece con $\sqrt{\alpha}$ (más lento: 1, 1.41, 1.73, 2...).

La Intuición Brutal: Imagina el equipo de relevos otra vez. Un corredor puede tener un mal día (tardar mucho) o un buen día (tardar poco). Es muy aleatorio.

Si tienes 100 corredores: Es imposible que los 100 tengan un mal día a la vez. Es imposible que los 100 tengan un buen día a la vez. Lo normal es que el corredor rápido compense al lento.

El Efecto de “Familia”: A medida que α crece, la suerte se cancela a sí misma. La campana se ensancha (la varianza absoluta crece, sí), pero se ensancha mucho más despacio de lo que se mueve hacia la derecha. Relativamente, la campana se vuelve más afilada.

Conexión Gamma(4) vs Gamma(5): Esto es la Propiedad Reproductiva. Si $X \sim \text{Gamma}(4, \lambda)$ y $Y \sim \text{Gamma}(1, \lambda)$ (una exponencial), entonces:

$$X + Y \sim \text{Gamma}(5, \lambda)$$

Esto significa que la Gamma no es una curva estática. Es una película. Empiezas con una Exponencial ($\alpha = 1$). Incertidumbre máxima. Sumas otra ($\alpha = 2$). La media se duplica, pero la incertidumbre no se duplica (solo $\times 1.4$). La curva empieza a formarse. Sumas otra... Cuando llegas a $\alpha = 100$, la curva es una campana perfecta y estrecha (relativamente).

Conclusión Brutal: La varianza te dice que acumular pasos estabiliza el sistema. Cuantos más pasos tiene una tarea compleja (α alto), más predecible es el tiempo total, porque los extremos estadísticos se diluyen.

Resumen Visual

Concepto	Fórmula	Significado “Brutal”
Complejidad	$t^{\alpha-1}$	Cada paso añade una Dimensión al universo de posibilidades.
Media	α / λ	El tiempo se apila como ladrillos. 4 ladrillos miden 4 veces más.
Varianza	α / λ^2	Los errores se cancelan. El caos individual muere en el colectivo.

Esta propiedad no afecta tanto a la forma de la curva pequeña, sino a lo que pasa cuando x es gigante.

4. La Aproximación de Stirling: “Cuando lo Discreto se vuelve Continuo”

$$\Gamma(n+1) \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

¿Qué es? Es una forma de calcular $100!$ sin multiplicar 100 veces.

Intuición Brutal (Entropía): Imagina un gas con 10^{23} moléculas. Calcular sus permutaciones exactas es imposible. Stirling nos dice que, cuando los números son enormes, el factorial se comporta como una función suave y predecible. En Física Estadística, usamos esto para derivar la Entropía ($S = k \ln \Omega$). El término $\ln(n!) \approx n \ln n - n$ es la base de toda la termodinámica moderna. Es la propiedad que dice que “el todo es más simple que la suma de las partes” cuando hay muchas partes.

Demostración (Método de Laplace) Vamos a demostrar Stirling directamente desde la definición de la integral Gamma. No va a ser una “demostración de servilleta”. Vamos a usar el Método de Laplace, que es la técnica definitiva para aproximar integrales que tienen un pico muy alto. Prepárate, porque vas a ver cómo una integral monstruosa se convierte en una simple Campana de Gauss.

PASO 1: El Campo de Batalla Partimos de la definición exacta para un entero n :

$$\Gamma(n+1) = n! = \int_0^\infty x^n e^{-x} dx$$

El problema es que x^n crece brutalmente y e^{-x} decrece brutalmente. Pelean a muerte.

Para ver quién gana y dónde ocurre la batalla, pasamos todo al exponente usando la identidad $A = e^{\ln A}$.

$$x^n = e^{n \ln x}$$

$$\Gamma(n+1) = \int_0^\infty e^{n \ln x - x} dx$$

Llamemos a la función del exponente $f(x)$:

$$f(x) = n \ln x - x$$

PASO 2: Encontrar la Cima (El Pico) La función $e^{f(x)}$ tendrá su máximo donde $f(x)$ tenga su máximo. Derivamos e igualamos a 0:

$$f'(x) = \frac{n}{x} - 1 = 0 \implies x = n$$

Intuición Brutal: La “batalla” ocurre casi enteramente alrededor de $x = n$. Si $n = 100$, el pico de la función está en 100. Lejos de 100, la función es casi cero. Por tanto, solo nos importa cómo se comporta la curva muy cerca de n .

PASO 3: La Cirugía Local (Taylor) Vamos a aproximar la función $f(x)$ alrededor del pico $x = n$ usando una Serie de Taylor de segundo orden (parábola).

$$f(x) \approx f(n) + f'(n)(x-n) + \frac{1}{2}f''(n)(x-n)^2$$

Calculemos los trozos: Altura ($f(n)$): $n \ln n - n$. Pendiente ($f'(n)$): Es 0 (porque es un máximo). Curvatura ($f''(n)$): La derivada de $\frac{n}{x} - 1$ es $-\frac{n}{x^2}$. Evaluada en n : $f''(n) = -\frac{n}{n^2} = -\frac{1}{n}$.

Sustituyendo en Taylor:

$$f(x) \approx (n \ln n - n) - \frac{1}{2n}(x-n)^2$$

PASO 4: La Transformación (Reconstruyendo la Integral) Ahora volvemos a meter esto en la integral original $\int e^{f(x)} dx$:

$$\Gamma(n+1) \approx \int_0^\infty e^{\left[n \ln n - n - \frac{(x-n)^2}{2n}\right]} dx$$

Usamos la propiedad de las potencias ($e^{A+B} = e^A \cdot e^B$) para sacar lo que no depende de x (las constantes) FUERA de la integral.

$$\Gamma(n+1) \approx \underbrace{e^{n \ln n - n}}_{\text{Constante}} \cdot \int_0^\infty e^{-\frac{(x-n)^2}{2n}} dx$$

Simplifiquemos la constante:

$$e^{n \ln n - n} = e^{\ln(n^n)} \cdot e^{-n} = n^n e^{-n}$$

¡Mira eso! Ya tenemos la mitad de la fórmula de Stirling ($\frac{n^n}{e^n}$). Ahora solo nos falta encontrar el $\sqrt{2\pi n}$.

PASO 5: La Aparición de Gauss Nos queda resolver la integral:

$$I = \int_0^{\infty} e^{-\frac{(x-n)^2}{2n}} dx$$

Hacemos un cambio de variable para centrarla. Sea $u = x - n$ (por tanto $dx = du$). Los límites cambian: de $0 \rightarrow \infty$ pasa a ser $-n \rightarrow \infty$. Pero como n es enorme (asumimos $n \rightarrow \infty$), podemos decir que integramos desde $-\infty$ a $+\infty$ porque la cola izquierda es despreciable.

$$I \approx \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2n}} du$$

¿Te suena esto? Es una Integral Gaussiana del tipo $\int e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$. En nuestro caso, $a = \frac{1}{2n}$.

$$I = \sqrt{\frac{\pi}{1/(2n)}} = \sqrt{2\pi n}$$

PASO 6: El Ensamblaje Final Juntamos la Constante (Paso 4) con la Integral (Paso 5):

$$\Gamma(n+1) \approx (n^n e^{-n}) \cdot (\sqrt{2\pi n})$$

Reordenando términos:

$$\Gamma(n+1) \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

Q.E.D. (Quod Erat Demonstrandum)

¿Dónde está la “Intuición Brutal” aquí? Fíjate en lo que acabamos de hacer. Hemos descompuesto el factorial en dos partes geométricas:

La Altura del Pico ($\frac{n^n}{e^n}$): Es el valor máximo de la función $x^n e^{-x}$. Representa la “moda” de la probabilidad.

La Anchura del Pico ($\sqrt{2\pi n}$): Viene de la integral gaussiana. Representa la “incertidumbre” o la dispersión alrededor del máximo.

Conclusión: El factorial de un número gigante ($n!$) es simplemente el Volumen de una Campana de Gauss que tiene:

Altura descomunal: $(n/e)^n$.

Anchura creciente: $\sqrt{n}$.

Stirling no es una aproximación mágica; es la constatación de que cualquier función elevada a una potencia gigante y multiplicada por un decaimiento se convierte en una Gaussiana. Es el Teorema Central del Límite disfrazado de cálculo integral.

Cuando α (o n) pasa de ser mayor que 1 (la montaña) a ser menor que 1 (el tobogán infinito), la física del sistema cambia radicalmente.

El Cambio de Paradigma: $\alpha > 1$: El tiempo trabaja a favor del evento (construye el resultado). $\alpha < 1$: El tiempo trabaja en contra del evento (disipa la oportunidad).

Una Gamma con $\alpha < 1$ surge naturalmente si tienes una Exponencial cuya velocidad (λ) varía aleatoriamente. Imagina un servidor web que recibe peticiones. A veces va lento (pocas peticiones). A veces va a la velocidad de la luz (ataque DDoS, millones de peticiones). Si mezclas tiempos de espera rápidos y lentos, pero hay muchísimos más rápidos que lentos, obtienes una Gamma con $\alpha < 1$.

La Intuición: La singularidad en el cero ($+\infty$) te está diciendo: “Oye, lo más probable es que esto sea un evento instantáneo, un micro-ruido, un error fugaz”. Pero la cola larga (e^{-t}) te dice: “Pero si sobrevive al primer instante, puede durar mucho”.

Es la distribución de los Sistemas Rotos o Fragilizados. Si un material tiene micro-grietas ($\alpha < 1$), lo más probable es que se rompa nada más tocarlo ($t \approx 0$). Si un material está impoluto ($\alpha > 1$), lo más probable es que aguante un rato hasta que acumule fatiga.

Mortalidad infantil / filtro de calidad ($0 < \alpha < 1$) Aquí es donde la Gamma con $0 < \alpha < 1$ brilla. Lo que modela exactamente es el fenómeno de “MORTALIDAD INFANTIL” o “EL FILTRO DE CALIDAD”.

Aquí tienes 3 ejemplos brutales que no tienen nada que ver con chispas.

1. El Ejemplo de la “Bombilla Defectuosa” (Ingeniería)

Imagina una fábrica de bombillas baratas. El control de calidad es pésimo.

Hay dos tipos de bombillas:

- **Las Rotas de Fábrica (90%):** Tienen el filamento suelto. En cuanto les metes corriente, ¡PUM!, se funden en milisegundos.
- **Las Buenas (10%):** Si sobreviven al encendido, duran años.

La Distribución del Tiempo de Vida:

- $\alpha > 1$ (**Desgaste**): Una bombilla perfecta de Philips. Tarda 1000 horas en gastarse. La probabilidad de morir en $t = 0$ es nula.
- $\alpha < 1$ (**Defecto**): La bombilla barata. La inmensa mayoría muere en $t \approx 0$ (al encenderse). Pico infinito en el cero.

Si ves una bombilla que lleva encendida 1 hora, piensas: “¡Ah, esta es de las buenas!”. Su probabilidad de seguir viva aumenta cuanto más tiempo pasa.

Lo que modela: Un sistema donde el peligro máximo está en nacer. Si sobrevives al inicio, eres “inmortal” (o decaes muy lento).

2. El Ejemplo de la “Búsqueda en Google” (Informática)

Imagina que eres un servidor de Google midiendo el tiempo que tarda un usuario en hacer clic en un resultado.

- **El Usuario “Robot” (Impaciente):** Ve lo que quiere y hace clic en 0.1 segundos (la mayoría).
- **El Usuario “Abuelo” (Lento):** Lee, duda, se toma un café, y hace clic a los 20 segundos (muy pocos).

Si dibujas la curva de “tiempos de reacción”:

- Hay una montaña gigantesca de clics casi instantáneos ($t \rightarrow 0$).
- Hay una cola muy larga de gente lenta.

Si usaras una Exponencial ($\alpha = 1$), asumirías un ritmo constante. Si usas una Gamma con $\alpha = 0.5$, estás modelando un sistema dominado por reacciones reflejas instantáneas.

Lo que modela: Tareas que suelen ser automáticas/inmediatas, pero que a veces se atascan.

Vamos a construir el ejemplo más simple posible

El Escenario: La Lotería de las Cajas Imagina un supermercado con una política extraña. No eliges caja. Te asignan una al azar al entrar.

Hay dos tipos de cajeros:

- **Cajeros “Turbo” (90% del personal):** Son máquinas. Te cobran en 0.1 minutos de media ($\lambda_{fast} = 10$).
- **Cajeros “Zootropolis” (10% del personal):** Son perezosos. Tardan 10 minutos de media ($\lambda_{slow} = 0.1$).

Tú eres un cliente.

- Probabilidad de que te toque Turbo: $p = 0.9$.
- Probabilidad de que te toque Perezoso: $1 - p = 0.1$.

1. La Matemática (Suma de Densidades) La función de densidad de tu tiempo de espera total $f(t)$ es la suma ponderada de las dos exponenciales:

$$f(t) = 0.9 \cdot (\text{Exp}_A) + 0.1 \cdot (\text{Exp}_B)$$

Sustituyendo la fórmula $f(t) = \lambda e^{-\lambda t}$:

$$f(t) = \underbrace{0.9 \cdot (10e^{-10t})}_{\text{Parte Rápida}} + \underbrace{0.1 \cdot (0.1e^{-0.1t})}_{\text{Parte Lenta}}$$

Simplificando:

$$f(t) = 9e^{-10t} + 0.01e^{-0.1t}$$

2. La “Autopsia” de la Curva (¿Por qué sale la forma Gamma < 1?) Vamos a evaluar esta función en dos momentos clave. Aquí verás nacer la forma de “L” (singularidad + cola larga).

En el instante $t = 0$ (El Inicio):

- Parte Rápida: $9 \cdot e^0 = 9$. (¡GIGANTE!)
- Parte Lenta: $0.01 \cdot e^0 = 0.01$. (Insignificante).
- **Total:** ≈ 9.01 .

Intuición: La densidad empieza altísima porque el 90% de los clientes salen disparados casi al instante. Esto simula la singularidad vertical.

En el instante $t = 2$ (Un rato después):

- Parte Rápida: $9 \cdot e^{-20} \approx 0$. (Los clientes rápidos ya se han ido hace mucho).
- Parte Lenta: $0.01 \cdot e^{-0.2} \approx 0.008$. (Apenas ha bajado).
- **Total:** ≈ 0.008 .

Intuición: La curva ha colapsado desde 9 hasta casi 0. Pero no llega a 0 absoluto. Se queda flotando en una cola larga y plana.

Resultado Visual: Tienes una curva que empieza en el cielo (9), cae en picado verticalmente, y luego frena en seco y se vuelve horizontal. ¡Esa es exactamente la forma de una Gamma con $\alpha < 1$! ($\frac{1}{t^{0.5}}e^{-t}$).

¿Qué modela una Gamma con $0 < \alpha < 1$? Modela la **Incertidumbre sobre la Escala de Tiempo**. Cuando te enfrentas a un proceso Gamma(0.5), no sabes si estás en la cola rápida o en la lenta.

- En $t = 0$: Tienes miedo. Lo más probable es que sea un proceso rápido y termine ¡YA! (Por eso la densidad es infinita).
- Si pasa el tiempo ($t > 0$): Te das cuenta de que no te ha tocado la cola rápida. Suspiras. Ahora sabes que estás en la cola lenta. Tu probabilidad de terminar pronto cae en picado.
- **¿El modelo del supermercado de 2 cajeros es una Gamma?** NO. Es una “Mixtura Exponencial Finita” (o Hiper-exponencial). Es pariente, pero no es lo mismo.
- **¿Cuándo se convierte en Gamma?** Cuando la mixtura se vuelve infinita y continua.
- **¿Por qué usamos Gamma si la realidad es finita?** Porque la fórmula $\frac{1}{\sqrt{t}}e^{-t}$ es mucho más fácil de manejar matemáticamente que sumar una lista de 10,000 tipos de cajeros distintos. La singularidad es el precio que pagamos por la simplicidad del cálculo.

Tu intuición te dice: “Si tengo 2 tipos de cajeros, la densidad en 0 es alta, pero finita. Si tengo 1000, será muy alta, pero finita”. Correcto. Pero mira lo que pasa si organizamos los cajeros siguiendo una Ley de Potencia (Fractal).

Imagina una jerarquía de cajeros:

- Nivel 1: Cajeros normales ($\lambda = 1$). Hay pocos.
- Nivel 2: Cajeros rápidos ($\lambda = 10$). Hay muchos más.
- Nivel 3: Cajeros turbo ($\lambda = 100$). Hay muchísimos más.
- ...
- Nivel k : Cajeros luz ($\lambda = 10^k$). La cantidad de estos cajeros crece brutalmente.

Si sumas las densidades de todos estos niveles:

$$f(0) = \sum_{k=1}^{\infty} (\text{Peso}_k \cdot \lambda_k)$$

Si el “Peso” (probabilidad de que te toque ese cajero) no baja lo suficientemente rápido para compensar la velocidad λ_k (que tiende a infinito), la suma diverge:

$$f(0) \rightarrow \infty$$

Conclusión: La Gamma con $\alpha < 1$ es el límite matemático de tener una población donde los elementos “ultrarrápidos” son infinitamente abundantes en comparación con los lentos.

2. La Intuición “Granularidad” vs. “Fluido” Aquí está la respuesta profunda a tu pregunta: *¿No son una Gamma hasta que se hace con infinitos procesos?*

Respuesta: En el mundo real (discreto), NADA es una Gamma perfecta en el cero.

- En el mundo real, no hay tiempos infinitesimales. Hay un tiempo mínimo de Planck o un tiempo mínimo de reacción.
- Por tanto, en la realidad física, la curva siempre “chocará” contra el eje Y en un valor muy alto pero finito (el tiempo mínimo).

La Gamma es una **IDEALIZACIÓN CONTINUA**. Es como un círculo en la pantalla de tu ordenador.

- Si haces zoom, son píxeles cuadrados (suma finita). No es un círculo.
- Si te alejas, los píxeles son tan pequeños que asumes el límite continuo.

La distribución Gamma modela el comportamiento macroscópico de un sistema que, a nivel microscópico, tiene una estructura tan fina y rápida que, a efectos prácticos, se comporta como si tuviera una singularidad.

1.7.7 Gamma como normal al cuadrado

Estás intentando unir dos continentes que parecen estar en planetas distintos:

- El Continente Temporal (Tu intuición): Relojes, esperas, colas, impaciencia.
- El Continente Espacial (Mi deducción): Distancias, errores, dianas, geometría.

La pregunta del millón es: ¿Por qué el “Tiempo de Espera” se comporta igual que la “Distancia al Cuadrado”?

Aquí tienes el puente definitivo. No es matemático, es **TOPOLÓGICO**.

La “Compensación” es la pelea a muerte entre dos fuerzas opuestas:

1. El Plegado (Matemático): La transformación $t = r^2$. Esta fuerza intenta crear una singularidad (infinito) en el cero.
2. El Espacio Vital (Geométrico): La dimensión del universo. Esta fuerza intenta diluir la probabilidad a medida que te alejas del centro.

Vamos a ver la pelea asalto por asalto.

ASALTO 1: El Enemigo (La Fuerza del Plegado) Independientemente de la dimensión, siempre estamos midiendo “Energía” o “Radio al Cuadrado” ($t = r^2$). El efecto de esto sobre la densidad es siempre el mismo: el Jacobiano.

$$\frac{dr}{dt} = \frac{1}{2\sqrt{t}} \propto \frac{1}{r}$$

Lo que hace: Esta fuerza grita: “¡Multiplica la densidad por infinito cuando r se acerca a 0!”. ¿Por qué? Porque al elevar al cuadrado números pequeños ($0.1 \rightarrow 0.01$), comprimes el espacio brutalmente. Tienes que meter la misma probabilidad en un sitio mucho más pequeño, así que la densidad se dispara.

ASALTO 2: La Pelea en 1D (La Derrota) El Escenario: Estás en una línea recta. La Geometría: El “cascarón” a distancia r son solo dos puntos ($+r$ y $-r$). El Volumen: El tamaño de esos dos puntos es constante. No crece con r .

$$\text{Volumen} \propto r^0 = 1$$

La Batalla:

- Plegado: Tira hacia infinito ($\frac{1}{r}$).
- Geometría: Es constante (1).

Resultado:

$$\text{Densidad Final} \propto \frac{1}{r} \times 1 = \frac{1}{r} = \frac{1}{\sqrt{t}}$$

Gana el Plegado. La singularidad sobrevive. Por eso la Gamma(1/2) explota en el origen. Es la “claustrofobia” de 1D: no hay sitio para acomodar la compresión del cuadrado.

ASALTO 3: La Pelea en 2D (El Empate Técnico) El Escenario: Estás en un plano. La Geometría: El “cascarón” a distancia r es una Circunferencia. El Volumen: La longitud de la circunferencia es $2\pi r$. Crece linealmente.

$$\text{Volumen} \propto r^1$$

La Batalla:

- Plegado: Intenta concentrar la densidad ($\frac{1}{r}$).
- Geometría: Intenta esparcir la densidad porque cada vez hay más sitio (r).

Resultado (La Aniquilación):

$$\text{Densidad Final} \propto \frac{1}{r} \times r = \text{Constante}$$

Empate. Las fuerzas se cancelan exactamente. La singularidad $1/r$ muere a manos del crecimiento r . Por eso la Gamma(1) (Exponencial) empieza en un valor constante finito. Ni explota ni vale cero. Es el equilibrio perfecto.

ASALTO 4: La Pelea en 3D (La Victoria de la Geometría) El Escenario: Estás en el espacio. La Geometría: El “cascarón” a distancia r es una Esfera. El Volumen: La superficie de la esfera es $4\pi r^2$. Crece cuadráticamente.

$$\text{Volumen} \propto r^2$$

La Batalla:

- Plegado: Sigue intentando concentrar ($\frac{1}{r}$).
- Geometría: Esparce la densidad con mucha fuerza (r^2).

Resultado:

$$\text{Densidad Final} \propto \frac{1}{r} \times r^2 = r = \sqrt{t}$$

Gana la Geometría. Ahora la función vale CERO en el origen. Aunque el plegado intentó subir la densidad, hay tan poco volumen en el centro de una esfera 3D comparado con el exterior, que la probabilidad de estar en $r = 0$ se anula. Por eso la Gamma(3/2) (Maxwell-Boltzmann) empieza en cero y sube.

Lo que vamos a hacer es construir el edificio desde el ladrillo fundamental. El “ladrillo” es una Variable Normal Estándar ($Z \sim N(0, 1)$). Tu objetivo es entender por qué, si sumas cuadrados de estas variables, obtienes Gammas con pasos de $1/2$. Aquí tienes la deducción rigurosa (Transformación de Variable) mezclada con la intuición geométrica (El Plegado del Espacio).

1. EL ÁTOMO: 1 Dimensión ($Z^2 \sim \Gamma(1/2)$)

El Experimento: Tienes una partícula que oscila en una línea recta (1D) siguiendo una Normal Estándar.

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Queremos la distribución de su Energía (o radio cuadrado): $X = Z^2$.

La Deducción Rigurosa (Cambio de Variable): La variable X solo puede ser positiva ($x > 0$). Para un valor concreto de energía x , hay DOS posibles orígenes: $z = +\sqrt{x}$ y $z = -\sqrt{x}$. La conservación de la probabilidad dice:

$$f_X(x) dx = f_Z(\sqrt{x}) dz + f_Z(-\sqrt{x}) dz$$

Como la Normal es simétrica, ambos lados aportan lo mismo:

$$f_X(x) = 2 \cdot f_Z(\sqrt{x}) \cdot \left| \frac{dz}{dx} \right|$$

El Jacobiano (La derivada): Si $x = z^2 \implies z = \sqrt{x} = x^{1/2}$.

$$\frac{dz}{dx} = \frac{1}{2}x^{-1/2} = \frac{1}{2\sqrt{x}}$$

Sustitución Final:

$$f_X(x) = 2 \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-x/2} \right) \cdot \left(\frac{1}{2\sqrt{x}} \right)$$

Los 2 se cancelan. Agrupamos términos:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{x}} \cdot e^{-x/2}$$

$$f_X(x) = \frac{1}{\sqrt{\pi}\sqrt{2}} x^{-1/2} e^{-x/2}$$

¿Qué es esto? Compara esto con la fórmula general de una $\text{Gamma}(\alpha, \lambda)$:

$$f(x) \propto x^{\alpha-1} e^{-\lambda x}$$

- El exponente de la exponencial es $-x/2$. Por tanto, $\lambda = 1/2$.
- El exponente de la potencia es $-1/2$. Como $\alpha - 1 = -1/2$, entonces $\alpha = 1/2$.

CONCLUSIÓN 1: El cuadrado de una Normal Estándar es exactamente una $\text{Gamma}(1/2, 1/2)$.

- Intuición: Hemos “plegado” la línea recta por la mitad. Toda la probabilidad se ha comprimido contra el cero, creando la singularidad $x^{-1/2}$.

2. LA MOLÉCULA: 2 Dimensiones ($Z_1^2 + Z_2^2 \sim \Gamma(1)$)

El Experimento: Ahora estás en un plano 2D. Tienes dos coordenadas independientes Z_1, Z_2 . El radio cuadrado es $R^2 = Z_1^2 + Z_2^2$. Llamemos $X_1 = Z_1^2$ y $X_2 = Z_2^2$. Ambas son $\text{Gamma}(1/2)$. La Deducción Rigurosa (Propiedad Reproductiva): Sabemos que si sumas Gammas independientes con el mismo λ , los parámetros α se suman.

$$R^2 = X_1 + X_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda)$$

$$R^2 \sim \text{Gamma}(1/2 + 1/2, 1/2)$$

$$R^2 \sim \text{Gamma}(1, 1/2)$$

¿Qué es una $\text{Gamma}(1)$? Es una Exponencial.

$$f(r^2) \propto (r^2)^{1-1} e^{-r^2/2} = e^{-r^2/2}$$

CONCLUSIÓN 2 (El Círculo): En 2D, la singularidad desaparece ($x^0 = 1$). La densidad de energía empieza en un valor finito y decae. ¿Por qué? Porque al añadir una segunda dimensión, el “espacio” disponible para alejarse del centro ($2\pi r$) crece linealmente, compensando exactamente la compresión del plegado.

3. EL CRISTAL: n Dimensiones ($Z_1^2 + \dots + Z_n^2 \sim \Gamma(n/2)$)

El Experimento: Estás en un hiperespacio de n dimensiones. El radio cuadrado es la suma de n cuadrados independientes. La Deducción Rigurosa: Aplicamos la propiedad reproductiva n veces.

$$\alpha_{total} = \frac{1}{2} + \frac{1}{2} + \dots + \frac{1}{2} = \frac{n}{2}$$

$$R^2 \sim \text{Gamma}(n/2, 1/2)$$

Esto es lo que los estadísticos llaman la distribución Chi-Cuadrado con n grados de libertad (χ_n^2).

1.8 Distribución Beta

1.8.1 Introducción

1. INTRODUCCIÓN: “La Distribución de las Probabilidades”

¿Es importante?

Si la Normal es la Reina de la Estadística (errores, ruido), la Beta es la Reina de la Inferencia (conocimiento, predicción).

Gamma: Modela Cantidades (tiempo, distancia, energía). Dominio: $[0, +\infty)$.

Beta: Modela Proporciones (probabilidades, porcentajes, eficiencias). Dominio: $[0, 1]$.

(a) Teoría de la Probabilidad (Deductiva / Hacia Adelante)

Situación: Tú eres Dios. Tú has llenado la urna. Sabes PERFECTAMENTE que hay 70 bolas rojas y 30 azules ($p = 0.7$).

La Pregunta: “Si meto la mano y saco 10 bolas, ¿cuál es la probabilidad de sacar 8 rojas?”

El Flujo: Modelo $\rightarrow$ Datos.

Vas de las Causas (parámetros conocidos) a los Efectos (datos probables).

Es matemáticas pura. La respuesta es única y exacta.

(b) Inferencia Estadística (Inductiva / Hacia Atrás)

Situación: Tú eres un Humano. No sabes qué hay en la urna. Solo ves que has sacado 10 bolas y 8 eran rojas (*datos*).

La Pregunta: “Visto lo que he sacado... ¿cuántas bolas rojas diablos hay dentro de la urna?”

El Flujo: Datos $\rightarrow$ Modelo.

Vas de los Efectos (datos observados) a las Causas (parámetros desconocidos).

Es detectivesca. La respuesta nunca es 100% segura, es una estimación con incertidumbre.

¿Por qué la Beta es la Reina de la Inferencia?

Porque la Beta es la distribución que responde a la pregunta inversa.

Binomial (Probabilidad): Dado p , dime cuántos éxitos k saldrán.

Beta (Inferencia): Dado que salieron k éxitos, dime cuánto vale p .

La Intuición Brutal:

Volvamos a tu intuición de Gamma = Tiempo para completar tareas.

El Escenario: Eres un procesador (CPU) que trabaja a una velocidad aleatoria pero común (λ).

Te llegan dos encargos:

- Encargo A (α): “Procesa α archivos”. (Tu carga de trabajo es α).
- Encargo B (β): “Procesa β archivos”. (Tu carga de trabajo es β).

Como tu velocidad (λ) es la misma para ambos (es tu velocidad de reloj):

- El tiempo que tardas en A (X) es proporcional a la cantidad de archivos α .
- El tiempo que tardas en B (Y) es proporcional a la cantidad de archivos β .

La Pregunta de la Beta:

$$B = \frac{\text{Tiempo dedicado a A}}{\text{Tiempo Total}}$$

Como la velocidad es la misma, la proporción de TIEMPO es equivalente a la proporción

de CANTIDAD.

$$B \approx \frac{\text{Carga } \alpha}{\text{Carga } \alpha + \text{Carga } \beta}$$

3. ¿Por qué es una Variable Aleatoria y no un número fijo? Si $\text{Tiempo} \propto \text{Cantidad}$, entonces B debería ser exactamente $\frac{\alpha}{\alpha+\beta}$, ¿no?

Si $\alpha = 50$ y $\beta = 50$, la Beta debería ser 0.5 clavado.

¡NO! Aquí entra la magia de la Gamma (la varianza). Recuerda que la Gamma no es un tiempo fijo, es un tiempo estocástico (el proceso de Poisson tiene ruido).

Aunque tengas que procesar 50 archivos A y 50 archivos B, a veces tendrás “mala suerte” con los archivos A y tardarás un poco más de lo normal (X sube).

A veces tendrás “buena suerte” con los B y tardarás menos (Y baja).

Por tanto, la proporción de tiempo real $X/(X+Y)$ no será exactamente 0.5. Oscilará alrededor de 0.5. Esa oscilación es la Distribución Beta.

Cuanto más archivos tengas (α, β grandes), más se promedian los tiempos (Ley de Grandes Números), y la oscilación se reduce $\rightarrow$ La Beta se vuelve una aguja en 0.5.

Si tienes pocos archivos ($\alpha = 1, \beta = 1$), la varianza es brutal. Podrías tardar el doble en 1 archivo que en otro por puro azar. $\rightarrow$ La Beta es plana (incertidumbre total).

4. ¿Qué narices tiene esto que ver con la Inferencia Bayesiana? Ahora conectemos esto con las monedas y la “Reina de la Inferencia”.

El Experimento Mental: Tú no sabes si una moneda es justa. La lanzas y observas: 3 Caras (α), 7 Cruces (β). Tu cerebro intenta reconstruir la realidad (“La Moneda”).

El Salto Lógico: Para tu cerebro bayesiano, haber visto “3 Caras” es equivalente a haber tenido que esperar el tiempo necesario para que ocurran 3 eventos de un proceso de Poisson.

La “Intensidad” con la que la moneda produce caras se modela como el tiempo Gamma asociado a esas caras. La “Intensidad” con la que produce cruces se modela como el tiempo Gamma asociado a esas cruces.

La Probabilidad p de la moneda se define como: “De toda la ‘fuerza existencial’ de la moneda, ¿cuánta se va en producir caras?”.

O dicho en términos de tiempo: “De todo el tiempo que la moneda está ‘operando’, ¿cuánto tiempo está generando caras?”.

$$p = \frac{\text{Esfuerzo Caras}}{\text{Esfuerzo Total}} = \frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)$$

Resumen para tu intuición: Usamos la Gamma para modelar la “cantidad de realidad” (o evidencia) que hemos acumulado para cada opción. La Beta es simplemente la normalización de esas cantidades para que sumen 1. Es convertir Conteos Absolutos (Gamma) en Frecuencias Relativas (Beta).

La Beta es la distribución que modela “La personalidad de una moneda”.

Si te encuentras una moneda en la calle, no sabes si está trucada.

- ¿Es una moneda justa ($p = 0.5$)?
- ¿Está trucada hacia cara ($p = 0.9$)?
- ¿Está trucada hacia cruz ($p = 0.1$)?

La Distribución Beta representa tu creencia sobre cuánto vale ese p .

2. LA CONSTRUCCIÓN: Hija de dos Gammas

Me preguntas: ¿La Beta se deduce de la Gamma?

SÍ. ABSOLUTAMENTE. Y la forma de verlo es geométrica y preciosa.

Imagina que tienes dos cronómetros Gamma independientes:

- Reloj Azul ($X \sim \Gamma(\alpha)$): Mide el tiempo para completar α eventos “Azules”.
- Reloj Rojo ($Y \sim \Gamma(\beta)$): Mide el tiempo para completar β eventos “Rojos”.

Los pones en marcha a la vez. Paran en tiempos aleatorios X e Y .

Ahora te preguntas: “Del tiempo total que han estado corriendo los relojes, ¿qué fracción corresponde al Reloj Azul?”

$$B = \frac{X}{X + Y}$$

Esta variable B sigue una distribución Beta(α, β).

La Intuición Geométrica (El Corte del Pastel):

Imagina una tarta lineal de longitud total $X + Y$.

Cortas la tarta en el punto X .

La Beta es la posición relativa del corte.

- Si el Reloj Azul es rápido (Gamma pequeña) y el Rojo lento, el corte está cerca del 0.
- Si ambos son iguales, el corte oscila alrededor del 0.5.

Conexión Profunda:

La Beta es la proyección angular de dos magnitudes Gamma ortogonales. Si dibujas (X, Y) en un plano 2D, la Beta te dice la “pendiente” o el ángulo del vector resultante.

3. LA CONEXIÓN CON EL RESTO (El Mapa Familiar)

La Beta es el “Eslabón Perdido” que conecta lo discreto (Binomial) con lo continuo (Uniforme/Normal).

(a) Con la Uniforme (El Origen de la Ignorancia)

Si $\alpha = 1$ y $\beta = 1$:

Tienes Beta(1,1).

Significa: “He visto 0 caras y 0 cruces”. (Nota: Los parámetros suelen ser éxitos+1, luego lo veremos).

No tienes ni idea de cómo es la moneda. Todas las probabilidades son igual de posibles.

Resultado: La Beta se convierte en una Recta Plana (Distribución Uniforme).

(b) Con la Binomial (El Espejo Bayesiano)

Esta es la conexión más importante de la estadística (Inferencia Bayesiana).

Binomial: Tú conoces la moneda (p) y quieres adivinar los datos (k caras).

Pregunta: “Dado que la moneda es justa, ¿probabilidad de sacar 3 caras?”

Beta: Tú conoces los datos (k caras) y quieres adivinar la moneda (p).

Pregunta: “Dado que he sacado 3 caras y 1 cruz, ¿cuál es la probabilidad de que la moneda sea justa?”

Matemáticamente, son Conjugadas:

Si empiezas creyendo que la moneda es una Uniforme (Beta 1,1) y tiras la moneda:

- Sale CARA → Tu creencia se actualiza a Beta(2,1). (La curva se inclina a la derecha).
- Sale CRUZ → Tu creencia se actualiza a Beta(2,2). (La curva se centra).
- Sale CARA → Beta(3,2)...

La Beta es el registro histórico de la batalla entre éxitos (α) y fracasos (β).

(c) Con la Normal (El Límite)

Si tiras la moneda millones de veces ($\alpha \rightarrow \infty, \beta \rightarrow \infty$):

Sabes con mucha precisión que la moneda es, digamos, 0.5.

La curva Beta se vuelve una aguja muy fina y simétrica centrada en la media.

En el límite, la Beta se convierte en una Campana de Gauss (Normal) extremadamente estrecha.

Para que la variable $B = \frac{X}{X+Y}$ siga una distribución Beta(α, β) estándar, el parámetro de tasa (λ) o de escala (θ) TIENE QUE SER IDÉNTICO en ambas variables Gamma.

Si los λ son distintos, el resultado NO es una Beta. Es otra distribución mucho más fea (una Beta Generalizada o Hipergeométrica Confluente) que depende del cociente λ_1 / λ_2 .

Aquí tienes la demostración rigurosa de por qué se cancelan y por qué tu intuición de “proporción de carga” era la correcta.

La Demostración por Escala (La prueba del algodón) Sabemos por la Propiedad de Escalado de la Gamma que si $X \sim \Gamma(\alpha, \lambda)$, entonces podemos escribirla como:

$$X = \frac{1}{\lambda} \cdot G_{\alpha}$$

Donde $G_{\alpha} \sim \Gamma(\alpha, 1)$ es una Gamma “estándar” (canónica, sin unidades de tiempo).

Si tienes dos variables con el MISMO λ :

$$X = \frac{1}{\lambda} G_{\alpha}$$

$$Y = \frac{1}{\lambda} G_{\beta}$$

Sustituimos en la definición de la Beta:

$$B = \frac{X}{X+Y} = \frac{\frac{1}{\lambda} G_{\alpha}}{\frac{1}{\lambda} G_{\alpha} + \frac{1}{\lambda} G_{\beta}}$$

Sacas factor común $1/\lambda$:

$$B = \frac{\frac{1}{\lambda} G_{\alpha}}{\frac{1}{\lambda} (G_{\alpha} + G_{\beta})} = \frac{G_{\alpha}}{G_{\alpha} + G_{\beta}}$$

¡El λ desaparece!

La variable resultante B ya no depende de la velocidad del procesador, ni de las unidades de tiempo (segundos, horas). Solo depende de los parámetros de forma α y β .

Esto es una Beta pura.

¿Por qué es esto una “Probabilidad”?

Porque en estadística frecuentista, la probabilidad se define como “Frecuencia Relativa”.

$$p \approx \frac{\text{Número de Éxitos}}{\text{Número Total de Intentos}}$$

La Conexión Brutal: La variable Beta es la distribución del TIEMPO RELATIVO que has gastado en la tarea A. Y como el tiempo es proporcional a la cantidad (porque λ es constante), el tiempo relativo es un estimador directo de la CANTIDAD RELATIVA (o frecuencia) de eventos A en el total.

1.8.2 Otra forma de obtener la Beta

Estadísticos de Orden (Order Statistics)

Me preguntas si hay otra forma de llegar a la Beta. Sí. Y es quizás más intuitiva geométricamente que la de las Gammas. Se llama la Intuición de la Cuerda Anudada.

El Experimento Imagina una cuerda de longitud 1 metro.

Tiras al azar N nudos sobre la cuerda (Distribución Uniforme).

Te fijas en la posición del k -ésimo nudo (ordenándolos de izquierda a derecha).

Pregunta: ¿Dónde cae el nudo número k ?

No cae en un punto fijo. Cae en una zona probabilística.

Esa zona es una Distribución Beta.

$$X_{(k)} \sim \text{Beta}(k, N - k + 1)$$

O usando nuestra notación de éxitos/fracasos:

Si tiras $\alpha + \beta - 1$ nudos...

La posición del nudo número α sigue una distribución $\text{Beta}(\alpha, \beta)$.

La Demostración Visual “Brutal”

- **Beta(1,1) [Uniforme]:** Tiras 1 nudo ($N = 1, k = 1$). ¿Dónde cae? En cualquier sitio con igual probabilidad. Es la Uniforme.
Fórmula: $\text{Beta}(1, 1 + 1 - 1) = \text{Beta}(1, 1)$. Correcto.
- **Beta(2,1) [Triángulo]:** Tiras 2 nudos. Te quedas con el mayor ($k = 2$).
Para que el nudo mayor esté en 0.9, el otro nudo tiene que haber caído en el 0.9 de espacio a la izquierda. (Probabilidad alta).
Para que el nudo mayor esté en 0.1, el otro nudo tiene que haber caído en el 0.1 de espacio a la izquierda. (Probabilidad baja).
Resultado: Una rampa lineal hacia el 1. $f(x) = 2x$.
- **Beta(2,2) [Parábola]:** Tiras 3 nudos. Te quedas con el del medio (la mediana, $k = 2$).
Tienes que tener uno a la izquierda y otro a la derecha.
Eso es más probable que ocurra en el centro de la cuerda.
Resultado: Una campana simétrica.
¿Por qué es brutal esta visión?
Porque conecta la Beta con los Percentiles.
La Beta no es solo “proporción de mezclas” (visión Gamma).
La Beta es la distribución de “Dónde cae el percentil X si muestreamos al azar”.

1. La Metáfora: El “Efecto Mateo” o el “Rico se hace más Rico”

Imagina una urna que empieza con una bola blanca y una bola roja. En la probabilidad clásica (con reemplazo), si sacas una roja y la devuelves, la urna no cambia. El futuro no tiene memoria.

En la **Urnas de Pólya**, el universo premia el éxito: Si sacas una bola roja, no solo la devuelves, sino que añades c bolas rojas más del mismo color. Esto crea un bucle de retroalimentación: cuantas más rojas saques, más probable es que la próxima sea roja.

La intuición brutal: Es una metáfora de la vida. La primera bola es como una pequeña inversión o un primer éxito. Ese éxito atrae más recursos, lo que hace que el segundo éxito sea más probable. Es una dinámica de auto-refuerzo.

2. El Salto Filosófico: Del Caos al Orden (La Beta)

Aquí ocurre la magia geométrica y filosófica. Podrías pensar que, debido a este refuerzo, al final la urna terminará siendo 100% roja o 100% blanca. Pero no es así.

- **El Proceso:** A medida que el número de extracciones n tiende a infinito, la proporción de bolas de un color no se vuelve loca, sino que se estabiliza.
- **La Sorpresa:** Si repitiéramos este experimento de la urna (desde el inicio) en un millón de universos paralelos, la proporción final en cada universo sería distinta. En un universo se estabilizaría en 0.2, en otro en 0.85, en otro en 0.5.
- **La Conexión:** Si graficamos todas esas proporciones finales de nuestros universos paralelos, ¿qué dibujo obtenemos? **La Distribución Beta.**

Pensamiento filosófico: La Urna de Pólya demuestra que el azar inicial (las primeras bolas) determina el destino a largo plazo, pero ese destino se congela en un valor específico. Es una transición de una variable discreta (bolas) a una densidad continua (la curva Beta).

3. Conexión con la Física y la Geometría

Geométricamente, puedes visualizarlo como un punto moviéndose en un segmento $[0, 1]$.

- Cada bola roja empuja el punto hacia el 1.
- Cada bola blanca lo empuja hacia el 0.

Al principio, los saltos son grandes y violentos. A medida que la urna se llena de miles de bolas, la influencia de una sola bola es mínima. El sistema adquiere inercia.

Usaremos el **Esquema de la Urna de Pólya clásico** (con refuerzo $c = 1$ para simplificar, aunque funciona para cualquier c).

1. El Escenario Formal: La Dinámica

Imagina una urna con: α bolas rojas (éxito), β bolas blancas (fracaso) y un total inicial $N = \alpha + \beta$.

Regla de Evolución: Extraemos una bola al azar. Observamos su color y la devolvemos a la urna junto con 1 bola adicional del mismo color. Sea X_i una variable indicadora donde $X_i = 1$ si la i -ésima bola es roja, y $X_i = 0$ si es blanca.

2. El Principio de Intercambiabilidad (La Intuición Geométrica)

Primero, calculemos la probabilidad de una secuencia específica. Digamos que queremos sacar Rojo, Rojo, Blanco (R, R, B).

$$P(R_1, R_2, B_3) = \frac{\alpha}{\alpha + \beta} \cdot \frac{\alpha + 1}{\alpha + \beta + 1} \cdot \frac{\beta}{\alpha + \beta + 2}$$

Ahora, cambiemos el orden: Rojo, Blanco, Rojo (R, B, R).

$$P(R_1, B_2, R_3) = \frac{\alpha}{\alpha + \beta} \cdot \frac{\beta}{\alpha + \beta + 1} \cdot \frac{\alpha + 1}{\alpha + \beta + 2}$$

¡Mira los numeradores! En ambos casos tenemos $(\alpha \cdot (\alpha + 1) \cdot \beta)$.

¡Mira los denominadores! Son idénticos: el producto de los totales crecientes.

Conclusión Brutal (Intercambiabilidad): La probabilidad de cualquier secuencia no depende del orden, solo depende del número total de éxitos (k) y fracasos ($n - k$). Esto es crucial. Significa que, aunque el proceso tiene “memoria”, la “historia” exacta no importa para la probabilidad final, solo el acumulado.

3. La Probabilidad de k éxitos en n intentos

Si la secuencia específica no importa, la probabilidad de obtener exactamente k bolas rojas en n extracciones ($S_n = k$) es simplemente la probabilidad de una secuencia específica multiplicada por el número de formas de ordenar esos éxitos (el coeficiente binomial).

La probabilidad de una secuencia con k rojas y $n - k$ blancas es:

$$P(\text{secuencia}) = \frac{\alpha^{[k]} \cdot \beta^{[n-k]}}{(\alpha + \beta)^{[n]}}$$

Nota: Uso la notación del factorial ascendente (Pochhammer): $x^{[k]} = x(x+1)\dots(x+k-1)$.

Entonces, la probabilidad total es la **Distribución Polya-Eggenberger**:

$$P(S_n = k) = \binom{n}{k} \frac{\alpha^{[k]} \beta^{[n-k]}}{(\alpha + \beta)^{[n]}}$$

4. El Puente: De Factoriales a Funciones Gamma

Sabemos que el factorial ascendente se relaciona con la función Gamma (Γ) así: $x^{[k]} = \frac{\Gamma(x+k)}{\Gamma(x)}$. Sustituycamos esto en nuestra fórmula. Prepárate para una reestructuración masiva:

$$P(S_n = k) = \binom{n}{k} \frac{\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \cdot \frac{\Gamma(\beta+n-k)}{\Gamma(\beta)}}{\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\beta)}}$$

Reordenamos los términos para agrupar las constantes (α, β) y las variables (k, n):

$$P(S_n = k) = \binom{n}{k} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha + k)\Gamma(\beta + n - k)}{\Gamma(\alpha + \beta + n)}$$

¿Te suena lo que está en azul? Es el inverso de la función Beta: $\frac{1}{B(\alpha, \beta)}$.

5. El Truco del Integral (La Aparición del Fantasma)

Aquí es donde conectamos con la física. Usamos la definición integral de la función Beta para reescribir la parte derecha de la ecuación (la que tiene las k y n):

$$\frac{\Gamma(\alpha + k)\Gamma(\beta + n - k)}{\Gamma(\alpha + \beta + n)} = B(\alpha + k, \beta + n - k) = \int_0^1 \theta^{\alpha+k-1} (1-\theta)^{\beta+n-k-1} d\theta$$

Sustituycamos esta integral en nuestra ecuación de probabilidad original:

$$P(S_n = k) = \binom{n}{k} \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^{\alpha+k-1} (1-\theta)^{\beta+n-k-1} d\theta$$

Vamos a reorganizar los términos dentro de la integral para que veas lo que realmente está pasando. Agrupamos lo que tiene que ver con la Binomial y lo que tiene que ver con la Beta:

$$P(S_n = k) = \int_0^1 \underbrace{\left[\binom{n}{k} \theta^k (1-\theta)^{n-k} \right]}_{\text{Verosimilitud (Binomial)}} \cdot \underbrace{\left[\frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)} \right]}_{\text{Densidad a Priori (Beta)}} d\theta$$

6. La Interpretación Final (El Teorema de De Finetti)

¿Qué nos dice esta última ecuación? Nos dice que el proceso mecánico de la Urna de Pólya (sacar, devolver, añadir) es **matemáticamente indistinguible** del siguiente proceso de dos pasos:

1. **El Fantasma:** Primero, el universo elige un número secreto θ (la probabilidad de éxito) sacándolo de una distribución $\text{Beta}(\alpha, \beta)$.
2. **La Realidad:** Luego, usando ese θ fijo (que nadie ve), lanza una moneda n veces (Binomial).
La Deducción: La proporción de bolas rojas en la urna $M_n = S_n/n$ no converge a una constante determinista. Converge a una variable aleatoria Θ . La distribución de esa variable aleatoria límite Θ es exactamente la parte “A Priori” de nuestra integral:

$$\Theta \sim \text{Beta}(\alpha, \beta)$$

Resumen Visual: Empezaste con un proceso discreto, ruidoso y mecánico. Al llevarlo al límite, el ruido no desaparece, sino que se cristaliza en una forma suave y continua: la curva Beta. La Urna de Pólya es la **manifestación física** de tener una distribución Beta como “prior” en tu mente bayesiana.

1.8.3 Distribución polar probabilística

1. La Analogía del Pintor (Color vs. Cantidad)

Imagina que estás mezclando pintura.

Tienes un bote de Pintura Azul (X). La cantidad es aleatoria (sigue una Gamma α).

Tienes un bote de Pintura Roja (Y). La cantidad es aleatoria (sigue una Gamma β).

Vuelcas los dos botes en un cubo grande. Ahora tienes una mezcla.

Para describir esa mezcla, necesitas dos números totalmente distintos:

La Cantidad ($S = X + Y$): ¿Cuántos litros tienes en total?

Esto es la Suma de Gammas. Te dice si el cubo está lleno o vacío.

El Color ($B = \frac{X}{X+Y}$): ¿Qué tono de violeta ha quedado?

Si $B = 1$, es Azul puro.

Si $B = 0$, es Rojo puro.

Si $B = 0.5$, es Violeta perfecto.

ESTO ES LA BETA.

¿Por qué es una “Proyección Angular”?

Porque en un gráfico donde el eje X es “Litros de Azul” y el eje Y es “Litros de Rojo”:

La Cantidad (S) es cuán lejos estás del origen (la distancia radial).

El Color (B) es la DIRECCIÓN en la que miras.

Intuición Brutal:

Saber que tienes 100 litros de pintura (Magnitud Gamma) no te dice absolutamente nada sobre de qué color es (Proporción Beta).

Son variables independientes. La Beta es la distribución de los “colores posibles” que resultan de mezclar dos cantidades Gamma aleatorias.

2. La Visualización Geométrica: El Abanico

Vamos a verlo en el plano cartesiano paso a paso. Dibuja mentalmente:

El Plano: Eje Horizontal (X) es Azul. Eje Vertical (Y) es Rojo.

El Punto: Cae un punto aleatorio (x, y) en el plano. (Ej: 3 litros azules, 1 litro rojo).

El Rayo (La Beta): Traza una línea desde el origen $(0,0)$ que pase por ese punto.

Esa línea tiene una inclinación.

Todos los puntos sobre esa misma línea tienen el mismo color (la misma proporción 3:1), aunque tengan distintas cantidades.

La variable Beta es una etiqueta para cada uno de los infinitos rayos posibles del abanico.

Si el punto cae muy pegado al eje X (mucha X, poca Y), el ángulo es bajo $\rightarrow$ Beta alta (≈ 1).

Si el punto cae muy pegado al eje Y (poca X, mucha Y), el ángulo es alto $\rightarrow$ Beta baja (≈ 0).

Lo que te dice la Beta:

Si generas mil puntos aleatorios (X, Y) usando distribuciones Gamma, y trazas sus rayos hacia el origen... ¿dónde se acumulan esos rayos?

Si $\alpha > \beta$ (Más tendencia al azul), la mayoría de los rayos estarán tumbados cerca del eje X.

Si $\alpha = \beta$, los rayos se centran en la diagonal de 45 grados.

Distribución polar probabilística Lo que has intuitido se llama formalmente la Descomposición Radial-Angular de vectores aleatorios independientes.

Si tienes dos variables independientes X, Y (digamos, normales o gammas relacionadas), puedes describir el sistema de dos formas equivalentes:

- Coordenadas Cartesianas (Independientes): X e Y . Aquí ves dos Gammas separadas.
- Coordenadas Polares (Independientes): Radio (R) y Ángulo (Θ). Aquí ocurre la magia: El Radio es Gamma. El Ángulo es Beta.

$$X + iY = R \cdot e^{i\Theta}$$

R (Módulo): Mide la “cantidad total” ($X + Y$). Sigue una distribución Gamma. (Se asocia a e^{-r} , decaimiento real).

Θ (Argumento): Mide la “proporción interna” ($X/(X + Y)$). Sigue una distribución relacionada con la Beta. (Se asocia a $e^{i\theta}$, rotación).

Tu intuición: “Gamma $\leftrightarrow$ Exponencial Real” y “Beta $\leftrightarrow$ Exponencial Imaginaria”.

La realidad matemática:

- La Gamma controla la dilatación del vector (escalar real).
- La Beta controla la rotación del vector (grupo circular).

1.8.4 Deducción de Beta a partir de Gamma

EL ESCENARIO MATEMÁTICO

Sean X e Y dos variables aleatorias independientes:

$$X \sim \Gamma(\alpha, \lambda)$$

$$Y \sim \Gamma(\beta, \lambda)$$

Nota crucial: El parámetro de escala λ (rate) es idéntico.

Sus funciones de densidad (PDF) son:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0$$

$$f_Y(y) = \frac{\lambda^\beta}{\Gamma(\beta)} y^{\beta-1} e^{-\lambda y}, \quad y > 0$$

Como son independientes, la densidad conjunta es el producto:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} y^{\beta-1} e^{-\lambda(x+y)}$$

EL CAMBIO DE VARIABLE (El bisturí)

Queremos estudiar dos nuevas variables. Una que representa la Proporción (Beta) y otra la Magnitud Total.

Definimos la transformación $\Phi(x, y) \rightarrow (b, s)$:

- $B = \frac{X}{X+Y}$ (La proporción, nuestro objetivo).
- $S = X + Y$ (La suma, una variable auxiliar necesaria para el cambio).

Paso 1: Invertir el mapa Para hacer el cambio de variable, necesitamos expresar x e y en función de b y s .

Como $x + y = s$, entonces $B = x/s \implies x = s \cdot b$.

Como $y = s - x$, entonces $y = s - sb = s(1 - b)$.

$$x(b, s) = s \cdot b$$

$$y(b, s) = s \cdot (1 - b)$$

Dominio: $x, y > 0$ implica que $s > 0$ y $0 < b < 1$.

EL JACOBIANO (La distorsión del espacio)

Aquí está la clave geométrica que separa a los niños de los adultos. Al cambiar de coordenadas rectangulares (x, y) a estas coordenadas “proporcionales” (b, s) , el espacio se deforma. El Determinante Jacobiano $|J|$ mide cuánto se estira o encoge el volumen diferencial.

$$J = \det \begin{pmatrix} \frac{\partial x}{\partial b} & \frac{\partial x}{\partial s} \\ \frac{\partial y}{\partial b} & \frac{\partial y}{\partial s} \end{pmatrix}$$

Calculamos las derivadas parciales:

$$\frac{\partial x}{\partial b} = s ; \frac{\partial x}{\partial s} = b$$

$$\frac{\partial y}{\partial b} = -s ; \frac{\partial y}{\partial s} = 1 - b$$

$$J = \det \begin{pmatrix} s & b \\ -s & 1 - b \end{pmatrix} = s(1 - b) - (b)(-s) = s - sb + sb = s$$

INTUICIÓN PROFUNDA DEL JACOBIANO ($J = s$) ¿Por qué el factor de corrección es s ?

Imagina que estás cortando un pastel (Suma S) en proporciones (B).

Si el pastel es pequeño (s pequeño), una diferencia del 1% (db) es una migaja física minúscula.

Si el pastel es gigante (s grande), una diferencia del 1% (db) es un trozo enorme.

El espacio físico ($dx dy$) que corresponde a un intervalo de proporción (db) crece linealmente con el tamaño total (s).

LA FUSIÓN (Sustitución en la Conjunta)

La fórmula del cambio de variable para densidades es:

$$f_{B,S}(b,s) = f_{X,Y}(x(b,s), y(b,s)) \cdot |J|$$

Sustituimos todo en la ecuación gigante del principio:

$$f_{B,S}(b,s) = \left[\frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} (sb)^{\alpha-1} (s(1-b))^{\beta-1} e^{-\lambda(sb+s(1-b))} \right] \cdot s$$

Simplifiquemos el exponente: $sb + s - sb = s$. (La b desaparece del exponente).

Agrupemos los términos de potencias:

$$f_{B,S}(b,s) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{s^{\alpha-1} s^{\beta-1} \cdot s}_{\text{Términos de s}} \cdot \underbrace{b^{\alpha-1} (1-b)^{\beta-1}}_{\text{Términos de b}} e^{-\lambda s}$$

Fíjate en las potencias de s : $s^{\alpha-1} \cdot s^{\beta-1} \cdot s^1 = s^{\alpha+\beta-2+1} = s^{(\alpha+\beta)-1}$.

Reordenando para separar lo que es de s y lo que es de b :

$$f_{B,S}(b,s) = \underbrace{\left(\frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)} s^{(\alpha+\beta)-1} e^{-\lambda s} \right)}_{\text{Parte que depende de S}} \cdot \underbrace{\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} b^{\alpha-1} (1-b)^{\beta-1} \right)}_{\text{Parte que depende de B}}$$

(Nota: He multiplicado y dividido por $\Gamma(\alpha+\beta)$ para completar las formas conocidas).

EL MILAGRO DE LA FACTORIZACIÓN (Independencia)

¡Mira la ecuación anterior! La densidad conjunta se ha factorizado perfectamente en el producto de dos funciones independientes:

$$f_{B,S}(b,s) = g(s) \cdot h(b)$$

1. Analizando $g(s)$ (La Suma):

$$g(s) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)} s^{(\alpha+\beta)-1} e^{-\lambda s}$$

¡Esto es exactamente la definición de una distribución $\text{Gamma}(\alpha+\beta, \lambda)$!

Acabamos de demostrar rigurosamente la Propiedad Reproductiva: La suma de Gammas independientes es otra Gamma con la suma de parámetros.

2. Analizando $h(b)$ (La Proporción):

$$h(b) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} b^{\alpha-1} (1-b)^{\beta-1}$$

Esto es una función de densidad válida para $0 < b < 1$.

Y esto es, POR DEFINICIÓN, la Distribución Beta(α, β).

El término constante $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ es el inverso de la función Beta de Euler $1/B(\alpha, \beta)$, que actúa como factor de normalización.

CONCLUSIONES DE LA DEDUCCIÓN RIGUROSA

Hemos demostrado tres cosas brutales de un solo tiro:

1. **Génesis de la Beta:** La distribución Beta surge inevitablemente al mirar la proporción de una magnitud Gamma respecto al total. Fíjate que el λ ha desaparecido completamente de la parte $h(b)$. Esto confirma matemáticamente que la Beta es invariante a la escala temporal/física.
2. **Génesis de la Suma:** La magnitud total (S) sigue una Gamma sumada. El Jacobiano (s) fue vital aquí. Sin ese s extra que vino de la distorsión geométrica, la potencia de s habría sido $\alpha + \beta - 2$, y no habría encajado con la definición de Gamma($\alpha + \beta$).
3. **Independencia Estocástica:** Como $f(b, s) = f(b) \cdot f(s)$, las variables B y S son Independientes.

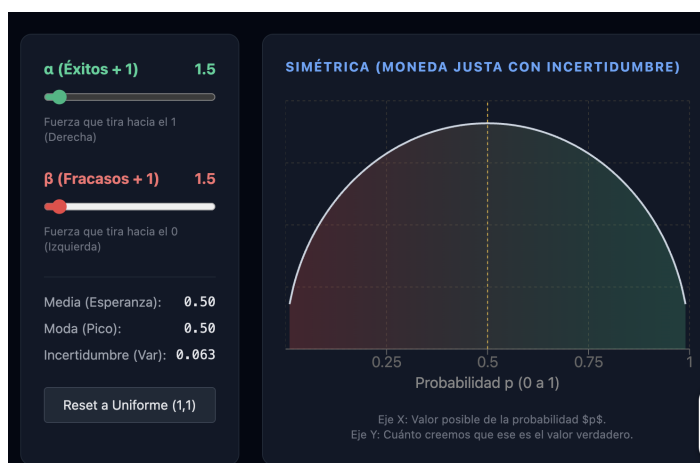
Intuición Física: Saber cuánto tardaste en total (S) no te da ni una pista sobre cómo repartiste el tiempo (B).

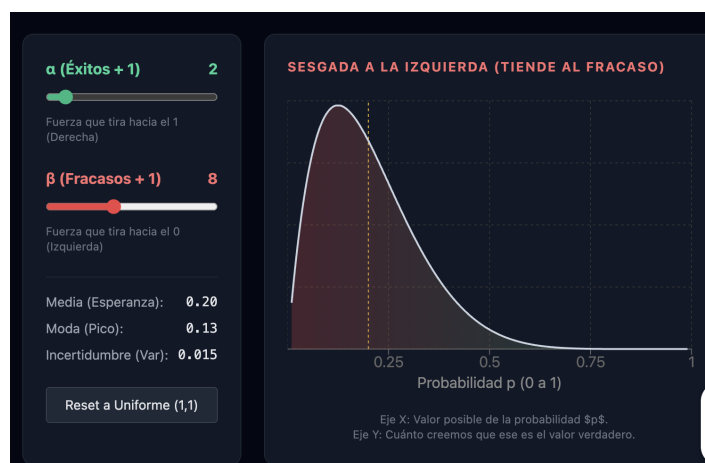
Intuición Geométrica: En coordenadas polares, la distancia al origen (R) no te dice nada sobre el ángulo (Θ).

Al factorizarse de esta forma, hemos demostrado rigurosamente que saber el volumen total de líquido que hay en el cubo (Suma S) no aporta absolutamente ninguna información sobre el color de la mezcla (Proporción B). **Son variables ortogonales e independientes.** La Beta nace inmaculada como la medida pura de la inclinación en ese mar estocástico.

1.8.5 Visualización gráfica: convexidad, colas e inercia









1. El Régimen del “Chispazo” o la Inestabilidad ($\alpha < 1, \beta < 1$)

Ejemplo: Beta(0.5, 0.5). La curva es una U (se pega a las paredes 0 y 1).

La Batalla en la Mezcla (La Beta): Tienes dos grifos espasmódicos.

- **Caso A:** El grifo Azul tose (casi 0) y el Rojo suelta algo.
Mezcla: Casi todo Rojo. $B \approx 0$.
- **Caso B:** El grifo Rojo tose (casi 0) y el Azul suelta algo.
Mezcla: Casi todo Azul. $B \approx 1$.
- **Caso C (El Centro):** Para que la mezcla sea violeta (0.5), ambos grifos tendrían que soltar *exactamente* la misma cantidad al mismo tiempo.

Como ambos tienden a apagarse (ir al 0), la probabilidad de que coincidan en un valor intermedio es bajísima. O gana el Azul por incomparecencia del Rojo, o viceversa.

2. El Régimen de la Estabilidad ($\alpha > 1, \beta > 1$)

Ejemplo: Beta(2, 2) o Beta(10, 10). La curva es una **Campana** (se acumula en el centro).

La Física: Los grifos ahora son estables. Tienen inercia. El grifo Azul garantiza, digamos, 10 litros $\pm$ un poco de ruido. El Rojo igual.

La Batalla en la Mezcla:

- **El Centro:** Como ambos grifos son fiables y siempre aportan caudal, la mezcla casi siempre será equilibrada (≈ 0.5).
- **Los Extremos:** Para que $B \approx 1$ (solo Azul), el grifo Rojo tendría que fallar y no echar nada.
¡Pero una Gamma(10) tiene probabilidad 0 de ser 0! (como vimos en “Gana la Geometría”).

Es imposible que falte pintura Roja. Por eso la curva Beta cae a 0 en los extremos. Cuanto

mayores sean α y β , más “chorro a presión” son los grifos, más determinista es la mezcla, y más estrecha es la campana (la varianza baja).

Lo que planteas sobre el **punto de inflexión** en la Beta tipo campana (ej: $\alpha = 10, \beta = 10$) es crucial. No es solo “una montaña”, es una estructura con zonas diferenciadas. Y lo más fuerte es que esos puntos de inflexión tienen un significado físico profundo conectado con la **inercia**.

1. El Mapa de la Montaña: Las Tres Zonas Imagina una Beta simétrica y “pesada” (ej: 10, 10). Tienes razón, no es toda cóncava.

- **Zona Externa (Las Colas): CONVEXA ($\cup$).**
Desde el 0 hasta subir un poco. Desde el 1 bajando hasta el final.
- **Zona Interna (La Cima): CÓNCAVA ($\cap$).**
Alrededor del pico (0.5).
- **Los Puntos de Inflexión:** El límite exacto donde pasas de “subir acelerando” a “subir frenando”.

2. La Intuición Gamma: ¿Por qué cambia la curvatura? Volvamos a tus dos grifos Gamma ($\alpha = 10, \beta = 10$) luchando. Recordemos la “batalla” descrita en el PDF: la Gamma es $t^{\alpha-1}$ (crecimiento/volumen) contra e^{-t} (decaimiento/probabilidad).

Zona Convexa (Las Colas): “El Vacío”

- Estás en los bordes (ej: $B = 0.1$). Aquí la probabilidad es bajísima.
- **La Física:** Para estar aquí, uno de los grifos (el Rojo) tiene que haber fallado estrepitosamente. Pero como $\alpha = 10$, el grifo tiene “fuerza”. Es muy difícil callarlo.
- **La Curvatura Positiva:** La curva se aplanas asintóticamente hacia el cero. Tiene que hacerlo suavemente (como un tobogán que se nivela al llegar al suelo) para no estrellarse. Esta suavidad al aterrizar es la convexidad. Representa la *resistencia absoluta* del sistema a tocar el cero perfecto.

El Punto de Inflexión: “El Muro de la Normalidad” Aquí está la magia. En una distribución Normal (y la Beta se parece a ella), los puntos de inflexión están exactamente en $\mu - \sigma$ y $\mu + \sigma$.

Conexión con tu PDF (Radio de Giro): El PDF define σ como el “Radio de Giro”.

La Intuición Brutal: El punto de inflexión es la frontera física del “núcleo” de la distribución.

- **Dentro de los puntos (Cóncava):** Es la zona de **Equilibrio**. Aquí, las fuerzas de los dos grifos se compensan. Estás “en la cima”. La gravedad te atrae al centro.
- **Fuera de los puntos (Convexa):** Es la zona de **Excepción**. Aquí ya no estás en equilibrio, estás cayendo por el precipicio de la improbabilidad.

El cambio de curvatura significa: “Hasta aquí es ‘normal’ (cóncavo, la cima). A partir de aquí es ‘raro’ (convexo, la cola).”

3. Visualizando la “Pelea de Fuerzas” (Derivada Segunda) Imagina que empujas una bola cuesta arriba por la curva de densidad.

- **Inicio (Cerca del 0):** La pendiente es positiva y cada vez es más empinada. (Convexa).
- **Fuerza:** La potencia $x^{\alpha-1}$ de la Gamma está empujando como loca hacia arriba. “¡Hay que salir del cero!”.

- **Punto de Inflexión:** De repente, la subida deja de acelerar y empieza a frenar.
- **Fuerza:** El término $(1 - x)^{\beta-1}$ (el otro grifo) empieza a gritar: “¡Eh, que te pasas! ¡Que yo también existo!”. La resistencia del otro componente empieza a ganar la batalla.
- **Cima (El 0.5):** La pendiente llega a 0. (Cónica).
- **Fuerza:** Empate técnico perfecto. Equilibrio de Nash entre los dos grifos.

Resumen “Brutal”

- **Beta(0.5, 0.5) [La U]:** Es pura convexidad. No tiene núcleo. Es todo “cola”. No hay estabilidad, solo extremismo.
- **Beta(10, 10) [La Campana]:** Tiene un núcleo cóncavo protegido por murallas convexas.
 - El núcleo cóncavo es la “masa” sólida del sistema (donde vive la verdad).
 - Las colas convexas son los “fleclos” donde vive el ruido.

El punto de inflexión es donde dejas de ver la “Señal” (el centro) y empiezas a ver solo “Ruido” (las colas). Es el límite de la inercia del sistema.

CASO 1: El “Martillo” o la Rampa Lineal

Parámetros: $\alpha = 2, \beta = 1$

Forma: Un triángulo rectángulo creciente ($y = 2x$). Una línea recta que sube.

- **La Física de los Grifos:**
 - Azul ($X \sim \text{Gamma}(2)$): Tiene inercia. Empieza en 0 y crece linealmente. Es un grifo que “coge fuerza”.
 - Rojo ($Y \sim \text{Gamma}(1)$): Es una Exponencial pura. Es el grifo “indiferente” o sin memoria. Suelta caudal constante probabilístico.
- **La Batalla:**
 - El Azul empuja con fuerza creciente.
 - El Rojo no opone resistencia estructurada (es plano/constante en su intención).
 - *Resultado:* La mezcla se inclina linealmente a favor del Azul.
- **Análisis de Curvatura:**
 - Curvatura Cero (Recta).
 - ¿Por qué? Porque el “empuje” de la Gamma(2) (x^1) se cancela exactamente con la indiferencia de la Gamma(1) (x^0). No hay aceleración ni frenazo. Es una “victoria por aplastamiento constante”.
- **Intuición de orden:** Tiras 2 nudos a la cuerda. Te quedas con el mayor. Es más probable que el máximo de dos números esté cerca del 1 que cerca del 0.

CASO 2: La “J” Invertida (El Sistema Roto vs. El Sano)

Parámetros: $\alpha = 0.5, \beta = 2$

Forma: Empieza en el infinito (eje Y), cae en picado, y luego muere suavemente en el 1.

- **La Física de los Grifos:**
 - Azul ($X \sim \text{Gamma}(0.5)$): Grifo roto/espasmódico. Tiende al chispazo (casi 0).
 - Rojo ($Y \sim \text{Gamma}(2)$): Grifo sano con inercia. Tarda un poco en arrancar, pero luego fluye bien.
- **La Batalla:**
 - El Azul grita “¡CERO!” con fuerza infinita al principio.
 - El Rojo intenta empujar hacia el centro, pero necesita tiempo.
 - *Resultado:* En la mezcla, casi siempre gana el “silencio” del Azul. La probabilidad se apila en $B \approx 0$.

- **Análisis de Curvatura:**

- Totalmente Convexa ($\cup$).
- Es un tobogán. La pendiente es siempre negativa.
- *Intuición:* Es la curva de la “Supervivencia Improbable”. Si sobrevives al chispazo inicial del azul (la zona de convexidad extrema), la probabilidad se aplanan, pero nunca llega a remontar (hacerse cóncava) porque el grifo Rojo es demasiado fuerte comparado con el Azul. Nunca hay equilibrio.

CASO 3: La “Cúpula” (El Primer Consenso)

Parámetros: $\alpha = 2, \beta = 2$

Forma: Una parábola perfecta simétrica ($\cap$).

- **La Física de los Grifos:**

- Ambos son Gamma(2). Ambos tienen “un poco de inercia” (lineal). Ninguno permite el cero absoluto.

- **La Batalla:**

- Ambos grifos odian el cero.
- Azul empuja lejos del 0. Rojo empuja lejos del 1.
- Se encuentran en el medio con fuerzas iguales.

- **Análisis de Curvatura:**

- Totalmente Cóncava ($\cap$).
- ¡Ojo aquí! A diferencia de la Normal (campana), la Beta(2,2) NO tiene puntos de inflexión. Es pura concavidad.
- Choca contra el suelo en 0 y 1 con un ángulo finito. No aterriza suavemente.
- *Intuición:* Es un “edificio” sin jardín alrededor. La zona de estabilidad (núcleo) ocupa todo el espacio. No hay “colas” convexas porque la inercia ($\alpha = 2$) es demasiado débil para generar asintotas suaves, pero suficiente para evitar el vacío.

CASO 4: La “Campana” (La Inercia Pesada)

Parámetros: $\alpha = 10, \beta = 10$

Forma: La clásica campana de Gauss (aproximadamente).

- **La Física de los Grifos:**

- Ambos son Gamma(10). Son “mangueras de bombero”. Chorros muy estables y potentes.
- El “Volumen de Historia” (t^9) es enorme.

- **La Batalla:**

- El Azul garantiza que habrá al menos un 40-60% de azul.
- El Rojo garantiza lo mismo.
- La mezcla es obligatoriamente 50/50 con muy poco margen de error.

- **Análisis de Curvatura (Aquí aparecen las 3 zonas):**

- **Zonas Convexas ($\cup$):** En los extremos (0 a 0.2 y 0.8 a 1). La probabilidad está “muerta”. La curva aterriza suavemente (asíntota) gracias a la alta potencia de las Gammas.
- **Zona Cóncava ($\cap$):** En el centro (0.2 a 0.8). El núcleo de estabilidad.
- **Puntos de Inflexión:** Son los muros de la fortaleza.

- **Diferencia Brutal con (2,2):**

- Beta(2,2) es una tienda de campaña (paredes tensas).
- Beta(10,10) es una montaña real (laderas suaves al principio, pico después). La aparición de la convexidad en las colas es la firma de la **Alta Estabilidad**.

CASO 5: El “Valle de la Muerte” (Polarización Suave)

Parámetros: $\alpha = 0.9, \beta = 0.9$

Forma: Una U suave.

- **La Física:**

- Estamos justo por debajo del límite de estabilidad ($\alpha = 1$).
- Los grifos “tosen” un poco, pero no están totalmente rotos como en el 0.1.

- **La Curvatura:**

- Totalmente Convexa ($\cup$).
- Pero fíjate: el centro no está vacío (como en 0.1, 0.1). Hay probabilidad en el 0.5, es solo que es el punto menos probable.
- *Intuición:* Es un sistema indeciso. Prefiere los extremos, pero “tolera” el centro. Es como una pelota en la cima de una colina muy suave; tiende a rodar, pero despacio.

Resumen de tu “Escáner de Curvatura”

1. **Miras la suma $\alpha + \beta$:**

- Si es gigante $\rightarrow$ Campana (Cóncavo en centro, Convexo en bordes).

2. **Miras los individuales:**

- Si alguno es $< 1 \rightarrow$ Singularidad (Pared vertical). Inestabilidad.
- Si es $= 1 \rightarrow$ Indiferencia (Línea recta o caída suave).
- Si es $= 2 \rightarrow$ Inercia Lineal (Triángulo o Parábola). Concavidad pura.
- Si es $> 2 \rightarrow$ Inercia Pesada (Aparición de colas convexas suaves).

La Beta Geométricamente: El “Mar” Gamma y el Abanico

Esta es la visualización definitiva que conecta las Gammas con la Beta. Olvida las fórmulas por un momento y piensa en topografía.

1. **El “Mar”: El Espacio Conjunto (X, Y)** Imagina un plano cartesiano (el suelo de una habitación).

- **Eje X:** Representa la variable $X \sim \text{Gamma}(\alpha)$ (ej. “Litros de pintura Azul”).
- **Eje Y:** Representa la variable $Y \sim \text{Gamma}(\beta)$ (ej. “Litros de pintura Roja”).

El “Mar” es la **Densidad Conjunta**. Como X e Y son independientes, la probabilidad de que ocurran a la vez es el **PRODUCTO** de sus densidades individuales:

$$f_{total}(x, y) = f_{Gamma}(x) \cdot f_{Gamma}(y)$$

Visualmente: Esto crea una “montaña” o nube de densidad sobre el plano.

- Si α, β son grandes, la montaña está lejos del origen (en el centro del mar).
- Si son pequeños, la montaña está pegada a la esquina $(0, 0)$.

2. **La Beta como Dirección (El Abanico)** Mira el “mar” (la montaña Gamma 3D) desde arriba. Cae una gota aleatoria (un punto de realidad) en la coordenada (x, y) .

- **El Radio ($S = X + Y$):** Es la distancia desde el origen hasta la gota. Te dice *cuánta* pintura hay. (Esto sigue otra distribución Gamma).
- **El Ángulo ($B = \frac{X}{X+Y}$):** Traza una línea recta (un rayo) desde el origen $(0, 0)$ que pase por la gota. La inclinación de esa línea es la Beta.

La Beta es el “Abanico”: La distribución Beta responde a la pregunta: “Si lanzas millones de puntos a este mar Gamma, ¿en qué direcciones se acumulan los rayos?”

- **Si $\alpha = \beta$ (Simetría):** La montaña Gamma es redonda (o simétrica respecto a la diagonal). La mayoría de los puntos caen en la diagonal $y = x$. Los rayos se acumulan en el centro ($45^\circ \approx \text{Beta } 0.5$).
- **Si $\alpha > \beta$ (Más Azul):** La montaña Gamma está estirada a lo largo del eje X. La mayoría de los puntos caen lejos en X y cerca de 0 en Y. Los rayos están “tumbados”. La Beta se acumula cerca del 1.

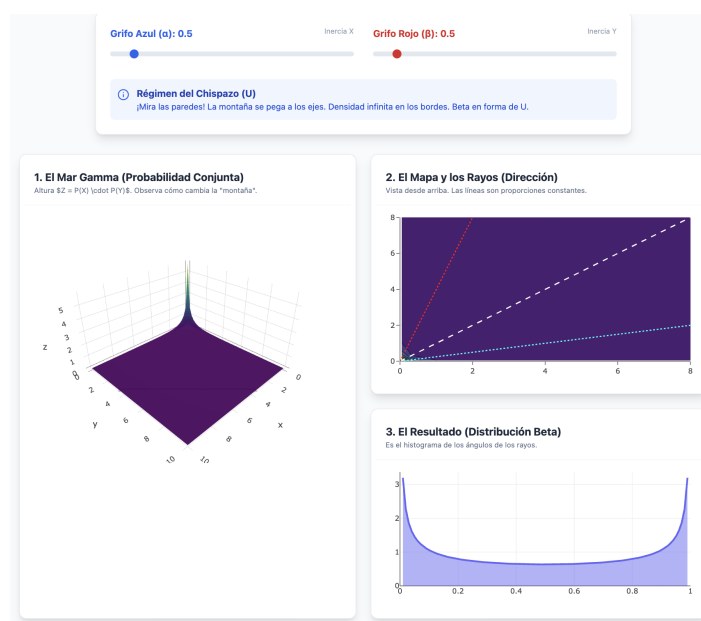
La Magia de la Independencia ¿Por qué es tan potente esta visión? Porque en este “Mar Gamma”, la distancia a la que estás (Suma) no te dice nada sobre la dirección en la que miras (Beta). **Son independientes.**

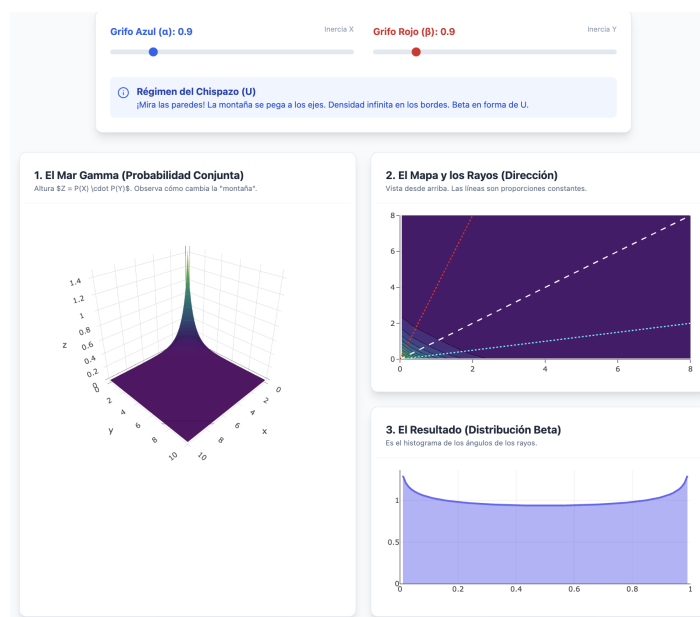
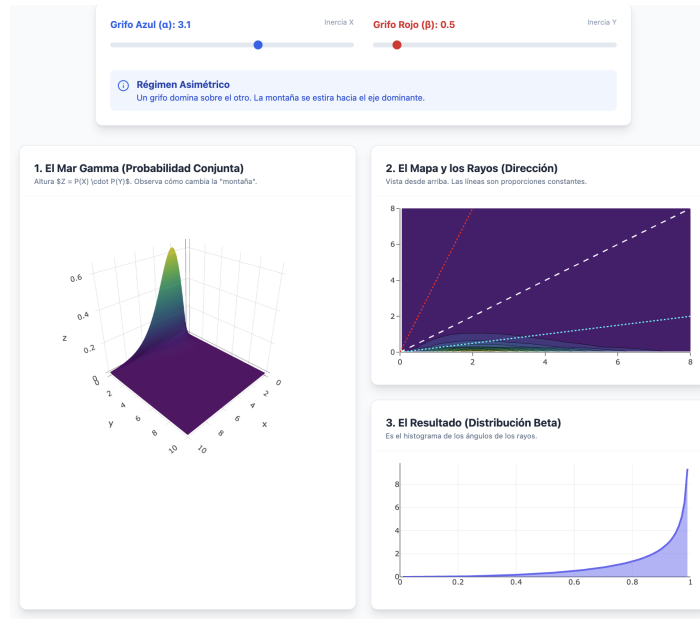
Puedes cortar la montaña cerca del origen (poca pintura) o lejos (mucha pintura), y la “forma” del corte transversal (el perfil de colores) es idéntica. Esa forma invariante es la **Distribución Beta**.

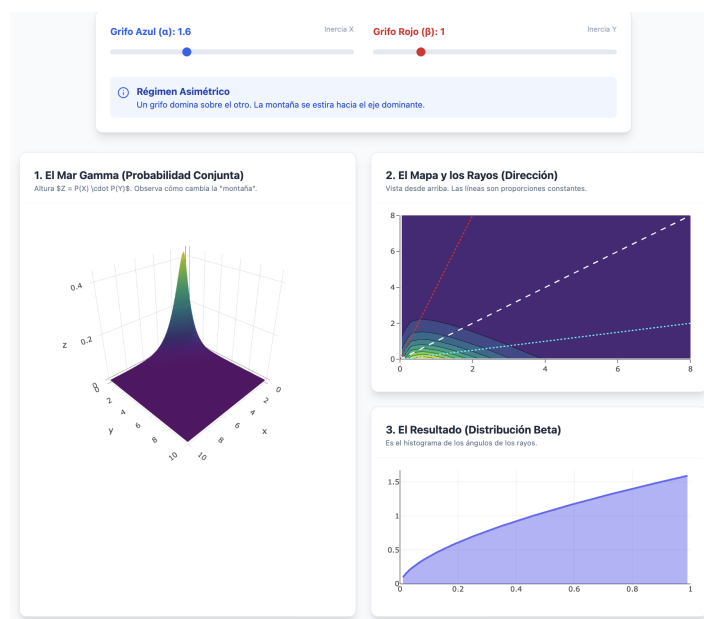
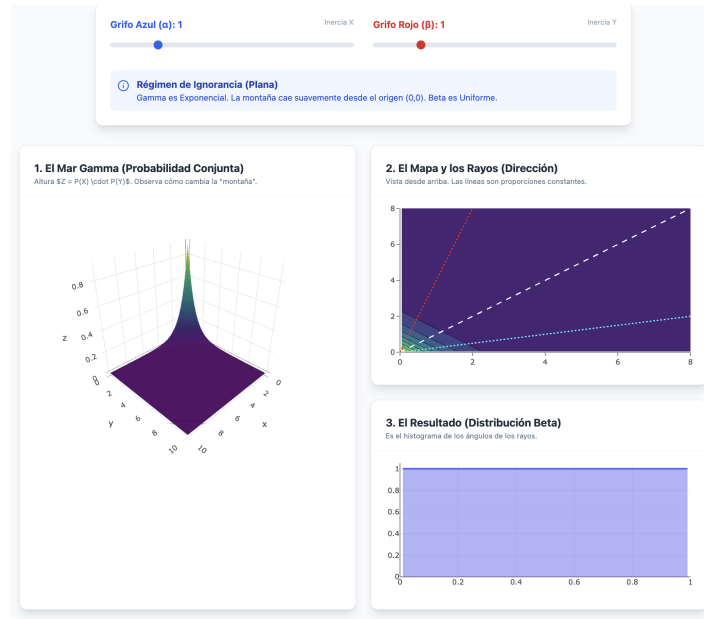
Nota sobre los Ejes (La Paradoja del Borde)

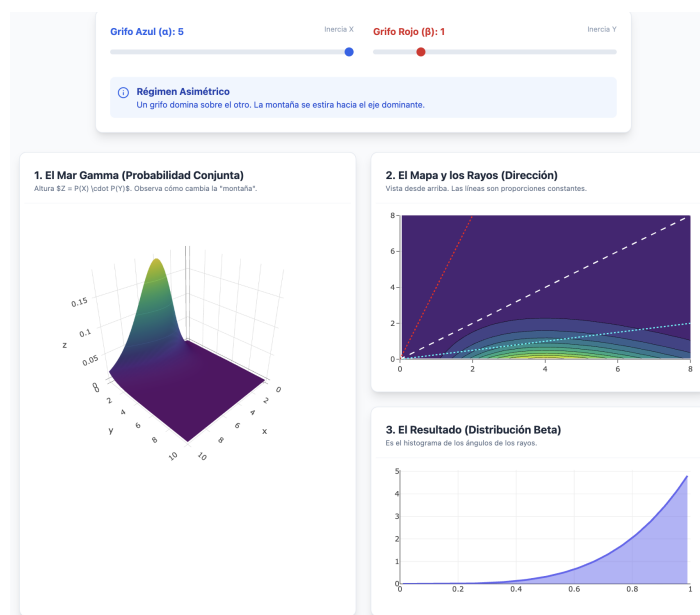
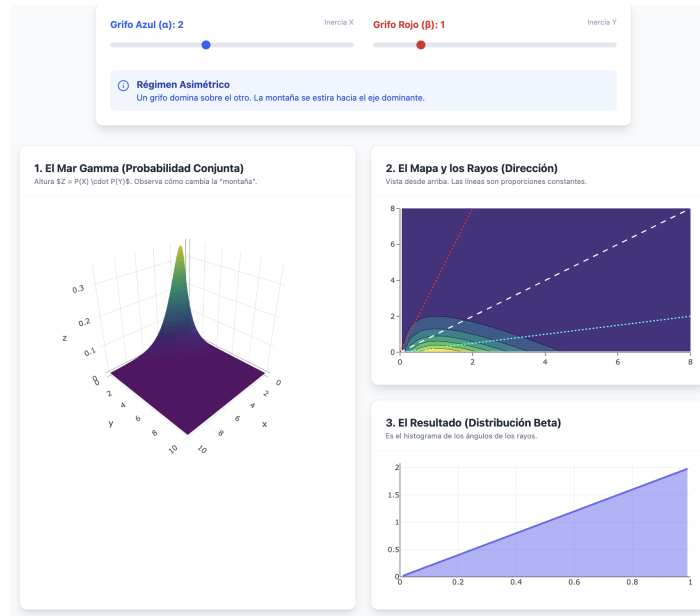
Es posible que te preguntes: *¿Cuál es la probabilidad en los ejes si $Y = 0$?* En los ejes no hay probabilidad, hay **cantidad**. La probabilidad es la **altura** (Z) de la montaña sobre esos ejes.

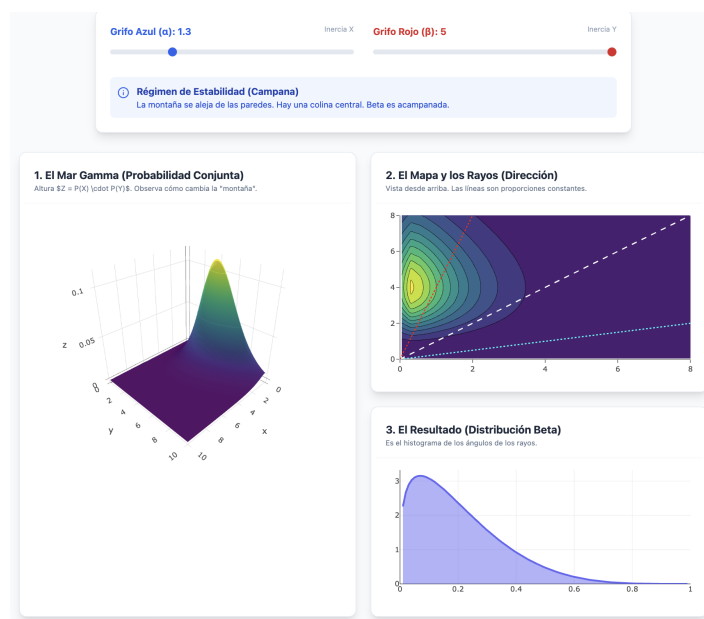
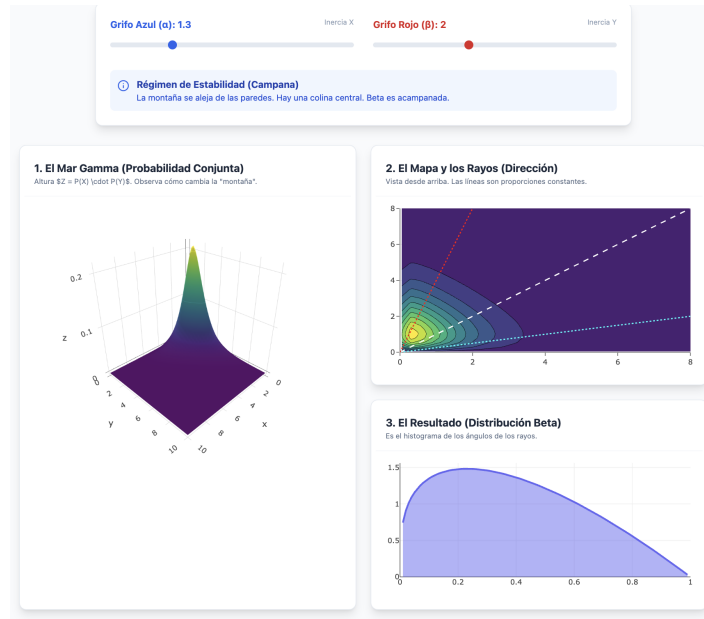
- **Caso $\alpha > 1$ (Playas):** La montaña nace suavemente del suelo. La altura en el eje es 0. La Beta ignora los extremos (forma de campana).
- **Caso $\alpha < 1$ (Acantilados):** La montaña tiene altura infinita en los ejes (singularidad). La “arena” de probabilidad se apila contra las paredes. La Beta es succionada por los extremos (forma de U).

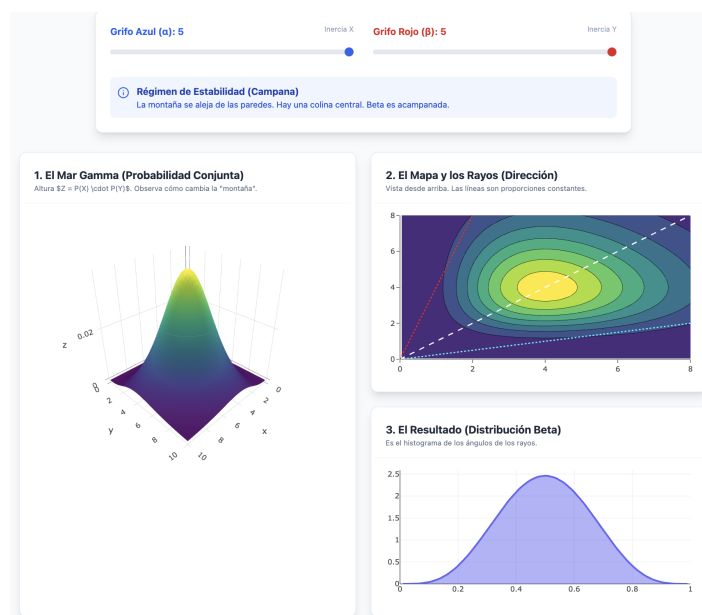
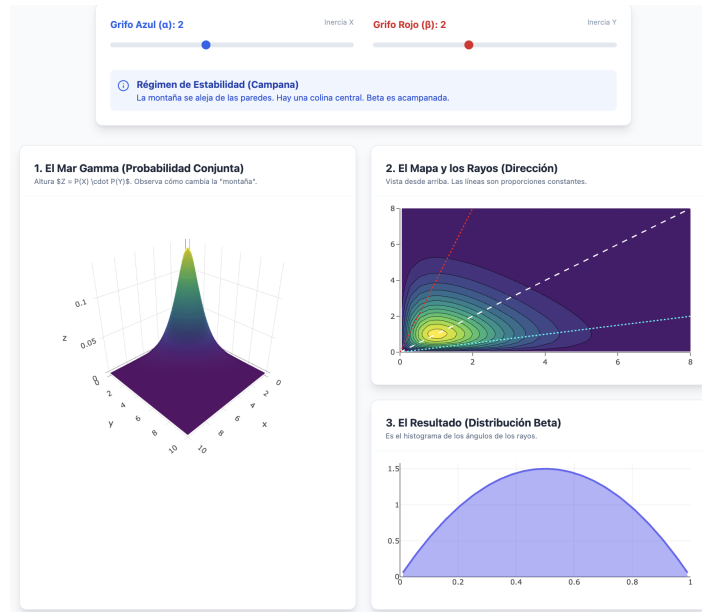


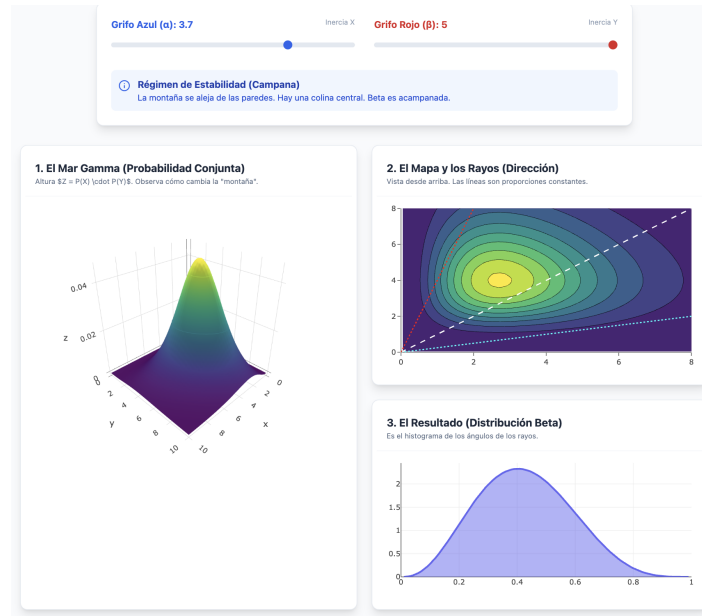












1.9 Intuición tras el Teorema Central del Límite

Si sumas muchas variables aleatorias independientes ($X_1 + X_2 + \dots + X_n$) con varianza finita, el resultado siempre converge a una Distribución Normal, sin importar la forma original de las X . ¿Por qué ocurre esto? (dos mecanismos intuitivos)

1. El mecanismo combinatorio (ejemplo: suma de Uniformes) La premisa: es mucho más fácil obtener un valor promedio que un valor extremo.

La geometría:

Suma de 1 (Uniforme): rectángulo (plano).

Suma de 2: triángulo (ya apunta al centro).

Suma de 3+: parábola/campana.

La razón: para obtener un extremo en la suma, todas las variables individuales deben conspirar para ser extremas a la vez (muy improbable). Para obtener un valor medio, hay infinitas combinaciones (una alta y una baja, dos medias, etc.). La combinatoria empuja la masa de probabilidad al centro.

2. El mecanismo de escala (ejemplo: Poisson) La premisa: al aumentar la magnitud de la suma, la granulosidad y los límites duros desaparecen.

La geometría:

Poisson pequeña (λ bajo): choca contra el muro del cero (asimétrica).

Poisson grande (λ alto): se aleja del muro (la restricción desaparece).

La razón (señal vs. ruido): el ruido (desviación típica) crece más lento ($\sqrt{n}$) que la señal (media, n).

Ruido relativo $\propto 1/\sqrt{n}$.

Al sumar mucho, las fluctuaciones individuales se vuelven insignificantes comparadas con el total, creando una curva suave y simétrica.

Conclusión La campana de Gauss no es una ley mágica: es la forma geométrica que adopta el caos cuando se acumulan muchos efectos pequeños, ningún efecto individual domina, y la energía (varianza) es finita.

CASO 1: La suma de Uniformes (Irwin–Hall) ¿Por qué si sumo "cajas planas" (Uniforme) me sale una "campana" (Normal)?

Paso 1: El experimento del dado (la versión discreta) Para entender la continua, miremos la discreta.

Imagina un dado de 6 caras.

1 dado: histograma plano. La probabilidad de sacar un 1 es igual que la de sacar un 6 (Uniforme).

Suma de 2 dados:

¿Para sacar un 2 (mínimo)? Solo hay una forma: (1+1).

¿Para sacar un 12 (máximo)? Solo hay una forma: (6+6).

¿Para sacar un 7 (medio)? ¡Hay un montón! (1+6, 2+5, 3+4, 4+3, 5+2, 6+1).

Resultado: un triángulo. Ya no es plano. El centro se ha hinchado porque hay más combinaciones para estar en el medio que en los extremos.

Paso 2: pasamos al continuo (variables $U(0,1)$) Ahora haz lo mismo con números reales entre 0 y 1.

Suma de 1 ($n = 1$): es una caja rectangular plana.

Suma de 2 ($n = 2$): es un triángulo.

Para que la suma sea casi 0, necesitas que ambos sean casi 0 (muy difícil, pocas combinaciones).

Para que la suma sea 1 (el medio), uno puede ser 0.1 y el otro 0.9, o 0.5 y 0.5... ¡Hay infinitas parejas que suman 1! (máxima probabilidad).

Suma de 3 ($n = 3$): aquí pasa la magia. El pico del triángulo se suaviza: se vuelve parabólico.

Para sacar un extremo (casi 0 o casi 3), necesitas tres milagros a la vez (tres números bajísimos o tres altísimos).

Para sacar el medio (1.5), te vale con que uno sea alto y los otros dos bajos, o los tres medios... La combinatoria "chupa" toda la probabilidad hacia el centro.

Paso 3: ¿por qué el truco de "sumar 12"? Esto es una leyenda de la programación antigua.

La varianza de una Uniforme $U(0,1)$ es $1/12$.

Si sumas 12 variables independientes, las varianzas se suman:

$$\text{Varianza total} = \frac{1}{12} + \frac{1}{12} + \dots \text{ (12 veces)} = 1.$$

¡Boom!

Al sumar 12, obtienes una distribución que:

- por el Teorema Central del Límite, ya es casi idéntica a una Normal (campana perfecta),
- tiene varianza = 1 (gratis, sin hacer raíces cuadradas),
- tiene media = 6 (porque la media de una es 0.5×12).

Así que, restándole 6 al resultado, los programadores antiguos generaban una Normal estándar $N(0,1)$ rapidísimo sin usar logaritmos ni cosenos.

CASO 2: La suma de Poisson ¿Por qué sumar eventos raros (asimetría) crea una campana simétrica?

Paso 1: el problema de la "pared" (asimetría) Imagina una Poisson con media baja ($\lambda = 1$ coche/minuto).

¿Puede venir -1 coche? No. Hay una pared dura en el cero.

¿Pueden venir 5 coches? Sí, aunque es raro.

Forma: la distribución está aplastada contra la pared izquierda y tiene una cola larga a la derecha. Es muy asimétrica.

Paso 2: alejándose de la pared (la suma) Ahora sumamos 50 horas de tráfico. Como suma de Poisson, el total también es Poisson, con parámetro

$$\lambda_{\text{total}} = 50.$$

Equivalente: $N \sim \text{Poisson}(50)$.

La media ahora está en 50.

La pared del cero sigue ahí, pero ahora está lejísimo (a 50 unidades de distancia).

La probabilidad de obtener 0 coches es e^{-50} (prácticamente cero absoluto).

Al alejarse de la pared, la distribución tiene espacio para "respirar" hacia la izquierda: puede fluctuar simétricamente (45 coches es tan probable como 55 coches).

Paso 3: ley de los grandes números (señal vs. ruido) En una Poisson, la desviación típica (el ruido) es la raíz de la media ($\sigma = \sqrt{\lambda}$).

Mira la proporción de ruido respecto a la señal (coeficiente de variación):

$$\frac{\text{Ruido}}{\text{Señal}} = \frac{\sqrt{\lambda}}{\lambda} = \frac{1}{\sqrt{\lambda}}.$$

Si $\lambda = 1$: el ruido es del 100% (caos total, asimetría brutal).

Si $\lambda = 100$: el ruido es $10/100 = 10\%$.

Si $\lambda = 1\,000\,000$: el ruido es 0.1% .

A medida que sumas (aumentas λ), el ruido relativo se hace insignificante. La "granulosidad" discreta de la Poisson se borra y lo que queda es una curva suave, simétrica y predecible alrededor de la media.

Resumen de la intuición Uniformes (suma): es combinatoria pura. Es infinitamente más fácil obtener un promedio mediocre (compensando altos con bajos) que obtener una suma extrema (donde todos tienen que ponerse de acuerdo para ser extremos). Esa acumulación en el medio es la campana.

Poisson (suma): es alejarse de la pared. Una Poisson pequeña es asimétrica porque choca contra el cero. Una Poisson grande vive tan lejos del cero que ya no siente la pared, y se relaja en una forma simétrica perfecta (Normal).

Imagina que estás midiendo el voltaje de una batería o la altura de una pared de ladrillos.

La señal (media, μ) Es la verdad: el valor que debería tener. Es la información útil.

Ejemplo: "Esta pared mide 3 metros".

El ruido (desviación típica, σ) Es la mentira: la basura, la imperfección, la fluctuación aleatoria alrededor de la verdad.

Ejemplo: "Bueno, un ladrillo sobresale 1 mm, otro está hundido 2 mm...".

El coeficiente de variación (CV) Es una forma de preguntar: "¿Cuánto molesta el ruido en comparación con el tamaño de la señal?"

$$CV = \frac{\text{Ruido}}{\text{Señal}} = \frac{\sigma}{\mu}.$$

Si CV es alto (p.ej. > 1): el ruido es más grande que la señal. Es el caos (una radio que solo suena "shhhhh").

Si CV es bajo (p.ej. < 0.01): el ruido es insignificante. La señal es nítida.

La magia del TCL: "el ruido se cancela, la señal se suma" Imagina que sumas n variables aleatorias (como sumar ladrillos para hacer una torre, o sumar electrones en un cable).

A. La señal se suma linealmente Si pones 100 ladrillos, la altura media se multiplica por 100:

$$\mu_{\text{total}} = n \cdot \mu.$$

B. El ruido se suma pitagóricamente Los errores son aleatorios: a veces un ladrillo es más grande (+1) y el siguiente es más pequeño (-1). Se compensan.

Por eso, el error acumulado no es la suma de errores: es la raíz cuadrada de la suma de varianzas (como la hipotenusa de un triángulo):

$$\sigma_{\text{total}} = \sqrt{n} \cdot \sigma.$$

Veredicto: ¿qué le pasa al CV?

$$CV_{\text{total}} = \frac{\sigma_{\text{total}}}{\mu_{\text{total}}} = \frac{\sqrt{n} \sigma}{n \mu} = \frac{\sigma}{\mu} \cdot \frac{1}{\sqrt{n}}.$$

El ruido relativo disminuye a medida que sumas más cosas (n).

Si sumas $n = 1$: tienes un ruido base.

Si sumas $n = 100$: el ruido relativo baja 10 veces.

Si sumas $n = 1\,000\,000$: el ruido relativo baja 1000 veces.

Conclusión: ¿por qué esto crea la Normal? En lo pequeño ($n = 1$, Poisson pequeña), el ruido es tan grande comparado con la señal que ves la granulosidad, la pared del cero y la asimetría: el CV es gigante.

En lo grande ($n \rightarrow \infty$), la señal ha crecido tanto que el ruido, aunque también ha crecido, se ha quedado pequeñito en comparación.

El TCL funciona porque la aleatoriedad se diluye a sí misma: al sumar millones de gotitas de agua (moléculas chocando), no ves el caos individual (ruido), ves un río fluyendo suavemente (señal).

Esa suavidad resultante, donde el ruido es solo una pequeña vibración simétrica alrededor de la media, se llama **distribución normal**.

Imagina que sumas n variables independientes idénticas ($X_1 + \dots + X_n$).

Corredor A: la media (la señal) La esperanza es un operador lineal:

$$\mathbb{E} [\sum X] = \sum \mathbb{E}[X] = n \cdot \mu.$$

Velocidad: crece proporcional a n .

(Si sumas 100 pilas de 1.5V, tienes 150V. Lineal.)

Corredor B: la varianza (la energía del ruido) La varianza (si son independientes) también es lineal:

$$\text{Var} (\sum X) = \sum \text{Var}(X) = n \cdot \sigma^2.$$

Velocidad: crece proporcional a n .

Corredor C: la desviación típica (el ruido real) La desviación es la raíz cuadrada de la varianza:

$$\sigma_{\text{total}} = \sqrt{\text{Var}_{\text{total}}} = \sqrt{n \cdot \sigma^2} = \sqrt{n} \cdot \sigma.$$

Velocidad: crece proporcional a $\sqrt{n}$.

La trampa: cuando el error es sistemático Imagina que el error no es aleatorio, sino sistemático:

"Mi regla está doblada y siempre mide 1 cm de más".

Error 1: +1

Error 2: +1

...

Error 100: +1

Aquí los errores no se cancelan, se apilan.

En ese caso, el ruido crece lineal:

$$\sigma_{\text{total}} = n \cdot \sigma.$$

Aquí la relación señal/ruido se mantiene constante. Nunca converge a una Normal.

Resumen El Teorema Central del Límite funciona gracias a la **independencia**. La independencia garantiza que los errores se peleen entre ellos (se cancelen parcialmente), frenando el crecimiento del ruido a $\sqrt{n}$, mientras dejan que la señal crezca libremente a n .

El Teorema Central del Límite (TCL) es la ley gravitatoria de la probabilidad: "Todo lo que se suma independientemente y tiene energía finita acaba cayendo en el pozo de la Normal".

GRUPO 1: Los ciudadanos ejemplares (convergen a la Normal) Estos cumplen el requisito sagrado: tienen varianza finita (energía acotada). Si sumas muchas copias de ellos, se convierten en una campana de Gauss.

La "Compensación" es la pelea a muerte entre dos fuerzas opuestas:

1. El Plegado (Matemático): La transformación $t = r^2$. Esta fuerza intenta crear una singularidad (infinito) en el cero.
2. El Espacio Vital (Geométrico): La dimensión del universo. Esta fuerza intenta diluir la probabilidad a medida que te alejas del centro.

1. **Bernoulli y Binomial.** La Binomial $B(n, p)$ ya es per se la suma de n Bernoullis. Si $n \rightarrow \infty$, la Binomial es Normal (De Moivre–Laplace).
2. **Geométrica y Binomial Negativa.** Ya vimos que la Binomial Negativa es la suma de r Geométricas. Si tu objetivo r es muy grande (buscas 1000 éxitos), la suma de esperas se vuelve simétrica y acampanada.
3. **Poisson.** La suma de Poissons independientes es otra Poisson con $\lambda_{\text{total}} = \lambda_1 + \dots + \lambda_n$. Cuando λ es muy grande, la Poisson se vuelve indistinguible de una Normal $N(\lambda, \lambda)$.
4. **Uniforme (continua y discreta).** Suma 12 variables Uniformes $U(0, 1)$: converge rapidísimo. Con 3 o 4 uniformes ya tienes una curva que parece una campana (Irwin–Hall).
5. **Exponencial y Gamma.** Suma de Exponenciales = Gamma. Cuando el parámetro de forma de la Gamma es grande, la asimetría desaparece y se convierte en Normal.

GRUPO 2: El impostor (casi, pero no) Hipergeométrica. El problema es que las tiradas no son independientes (sin reemplazo). Si la urna es gigantesca ($N \rightarrow \infty$) comparada con lo que sacas (n), se comporta como una Binomial y converge a la Normal; si la urna es pequeña, se queda deforme.

GRUPO 3: El anarquista (la distribución de Cauchy) Aquí está la madre del cordero: la Cauchy rompe el juego.

En tu PDF pone claramente: "No existen los momentos de orden $r \geq 1$... en particular no existe ni la media ni la varianza".

¿Por qué es devastadora? Tiene colas pesadas: es tan probable que salga un valor monstruosamente gigante que la varianza es infinita.

Si sumas 100 variables de Cauchy y haces el promedio... el resultado sigue siendo una Cauchy igual de ancha.

No se estrecha.

No se hace Normal.

No converge.

La Cauchy es el "agujero negro" de la probabilidad: si tu sistema tiene varianza infinita (energía infinita), el equilibrio de Gauss no existe. El caos nunca se calma.

Vamos a construir un pilar de 100 bloques.

Cada bloque tiene una altura que sigue una Distribución Normal $N(\mu, \sigma)$:

$\mu = 10$ cm (altura teórica).

$\sigma = 1$ cm (error de fabricación / tolerancia).

Queremos saber la altura total y el error total del pilar:

$$H_{\text{total}} = X_1 + X_2 + \dots + X_{100}.$$

CASO 1: Independencia (error aleatorio / ruido térmico) El escenario: compras los bloques a 100 proveedores distintos, o usas un molde diferente para cada uno.

El bloque 1 sale un poco alto (+0.5).

El bloque 2 sale un poco bajo (-0.4).

El bloque 3 sale perfecto (0.0).

La física: los errores son vectores que apuntan en direcciones aleatorias. Se pelean entre ellos. Hay interferencia destructiva.

La matemática: la varianza de la suma es la suma de las varianzas (porque la covarianza es 0):

$$\text{Var}(H_{\text{total}}) = \text{Var}(X_1) + \dots + \text{Var}(X_{100}) = 100 \cdot \sigma^2.$$

Pasamos a sigma (unidad real de longitud):

$$\sigma_{\text{total}} = \sqrt{100 \cdot \sigma^2} = 10 \cdot \sigma.$$

Has sumado 100 bloques, pero el error solo ha crecido 10 veces.

Intuición: el caos se disuelve.

CASO 2: Dependencia total (error sistemático / molde roto) El escenario: tienes un solo molde para fabricar los 100 bloques, pero el molde está defectuoso (dilatación térmica, o mal calibrado). Está sesgado.

Si el molde hace el primer bloque más grande ($+1\sigma$), hará todos los bloques más grandes.

Aquí las variables ya no son $X_1, X_2, \dots$ independientes. Físicamente, es la misma variable repetida 100 veces:

$$H_{\text{total}} = X_1 + X_1 + \dots + X_1 = 100 \cdot X_1.$$

La física: los errores son vectores que apuntan todos en la misma dirección. Se apilan. Hay resonancia constructiva total. No hay cancelación.

La matemática: sacamos constantes de la varianza ($\text{Var}(kX) = k^2 \text{Var}(X)$):

$$\text{Var}(100 \cdot X_1) = 100^2 \cdot \text{Var}(X_1) = 10\,000 \cdot \sigma^2.$$

Pasamos a sigma (unidad real de longitud):

$$\sigma_{\text{total}} = \sqrt{10\,000 \cdot \sigma^2} = 100 \cdot \sigma.$$

Mira la diferencia brutal:

Independiente: error total = 10 cm.

Dependiente: error total = 100 cm.

¿Por qué pasa esto? (el término cruzado) En tu PDF o apuntes tendrás esta fórmula general para la varianza de dos variables:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y).$$

En el caso 1 (independencia): $\text{Cov}(X, Y) = 0$. Los bloques no se hablan entre sí. El término cruzado desaparece.

En el caso 2 (dependencia total): Y es igual a X . La covarianza es máxima: $\text{Cov}(X, X) = \text{Var}(X)$.

No solo sumas las diagonales: sumas toda la matriz de varianzas-covarianzas llena de números positivos.

Ese término $2 \cdot \text{Cov}$ es la energía de interacción entre los bloques. Al ser dependientes, sus errores se potencian.

Resumen devastador $\sigma_{\text{total}} = \sqrt{n} \sigma$ ocurre cuando los errores se ignoran entre sí (matriz diagonal).

$\sigma_{\text{total}} = n \sigma$ ocurre cuando los errores conspiran juntos (matriz llena de unos).

En ingeniería, el primer caso es tolerancia (se asume aleatoriedad). El segundo caso es mala calibración (un desastre sistémico).

Existe una versión más potente del teorema (para procesos estocásticos y martingalas) que dice:

"Si las variables son dependientes, PERO la dependencia es débil y simétrica (intercambiable), y ninguna variable domina sobre las demás... entonces TAMBIÉN converge a una Normal."

La intuición brutal Imagina que la "dependencia" es una cuerda que ata a las bolas entre sí.

Dependencia fuerte: si saco la bola A, la bola B desaparece mágicamente. (Aquí no hay Normal que valga).

Dependencia de la hipergeométrica (débil): si saco la bola A, la probabilidad de B baja un poquito (de $1/N$ a $1/(N-1)$). Si N es grande, ese cambio es microscópico. La "cuerda" es tan fina que apenas se nota.

Por eso, visualmente, la hipergeométrica SÍ se convierte en una campana de Gauss.

La diferencia crucial: la "campana anoréxica" Aquí está el matiz que hace que todo encaje. La hipergeométrica converge a una Normal, PERO no a la misma Normal que la binomial.

Mira la varianza (la "gordura" de la campana):

Binomial (independiente):

$$\text{Var} = n p q$$

Hipergeométrica (dependiente):

$$\text{Var} = n p q \left(\frac{N-n}{N-1} \right)$$

Ese factor es el **factor de corrección de población finita (FPC)**.

Si n es pequeño comparado con N , el factor es casi 1. La "dependencia" es invisible. El TCL funciona igual que siempre.

Si n es grande (p.ej., $n = 0.8N$), el factor es pequeño (0.2). La varianza se desploma.

Visualización El TCL funciona, la forma es de campana, pero la dependencia (el hecho de que se acaben las bolas) aplasta la campana por los lados: la hace más estrecha, más precisa, menos aleatoria.

¿Cuándo se rompe todo? (la muerte de la gaussiana) El TCL "generalizado" falla en la hipergeométrica sobre todo en los bordes extremos, donde la dependencia se vuelve tiránica:

Cuando $n \approx N$: aquí la dependencia es total. Ya no hay aleatoriedad. La campana colapsa en una línea vertical (delta de Dirac).

Cuando p es minúsculo: aquí la asimetría es brutal. No hay suficientes éxitos para formar una campana. Se queda en una distribución torcida (como la Poisson).

1.10 Distribución de Cauchy

«¿Es la que utiliza Nassim Taleb en lugar de la normal para modelar colas largas?»

Sí. Nassim Nicholas Taleb, el autor de *El Cisne Negro*, divide el mundo en dos dominios: “Mediocristán” (donde manda la Campana de Gauss y los eventos extremos no importan) y “Extremistán” (donde mandan las distribuciones de colas pesadas o *fat tails*, y un solo evento extremo puede dominar el total).

La distribución de Cauchy es la pesadilla definitiva de la estadística clásica y el ejemplo favorito de Taleb para destruir los modelos financieros tradicionales. De hecho, la distribución de Cauchy es un caso especial de las distribuciones de Pareto o leyes de potencias (es una t-Student con exactamente 1 grado de libertad). Es la encarnación matemática del Caos.

1. La Trampa Visual: El lobo con piel de cordero

En tus apuntes te dan la función de densidad general: $f(x) = \frac{1}{\pi} \frac{b}{b^2 + (x-a)^2}$, y la estándar (con $a = 0, b = 1$): $f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$.

Si tú dibujas esta función, te va a salir una campana. Es simétrica, tiene un pico en el centro, parece inofensiva... Parece una distribución Normal.

Pero mira el denominador: $1 + x^2$.

En una distribución Normal, la función decae con e^{-x^2} . La exponencial aplasta las colas contra el suelo rapidísimo. La probabilidad de ver a un humano de 3 metros (un evento 6 sigma) es literalmente menor que los átomos del universo.

En la Cauchy, la función decae como $1/x^2$. **¡Eso es lentísimo!** Las colas se quedan “flotando” por encima del eje horizontal eternamente. Tienen tanta masa que los eventos extremos no son anomalías; son una certeza matemática.

2. La Propiedad Destructora: “No existe ni la media ni la varianza”

Tus apuntes sueltan la bomba más grande en la última línea: “No existen los momentos de orden $r \geq 1$ de una variable aleatoria Cauchy, en particular no existe ni la media ni la varianza.”

Intuición abismal: ¿Qué demonios significa físicamente que “no tenga media”?

Si tú coges una moneda, la lanzas 100 veces y apuntas la media de caras, te dará alrededor de 0.5. Si la lanzas 1.000.000 de veces, se clavará en 0.500000. La media muestral converge (Ley de los Grandes Números).

Si tú generas 100 números aleatorios de una Cauchy y haces la media, te dará, digamos, 3.4.

Generas 1.000 números, haces la media... y te da -8.2.

Generas 1.000.000 de números, haces la media... y te da 450.1.

¡La media nunca se estabiliza! ¿Por qué? Porque justo cuando crees que la media se está asentando, la “cola pesada” de la Cauchy te escupe un valor extremo de, por ejemplo, +10.000.000, que arrastra todo el promedio y lo destroza.

Matemáticamente, la integral de la esperanza $\int_{-\infty}^{\infty} x \frac{1}{\pi(1+x^2)} dx$ da $\infty - \infty$, lo cual es indefinido.

Por eso Taleb la adora: si los mercados financieros fueran Cauchy, hablar de “rendimiento medio” o “volatilidad (varianza)” es una aberración matemática. **¡No existen!**

3. El Engaño de los Parámetros a y b

Tus apuntes dicen que los parámetros son $a \in \mathbb{R}$ y $b > 0$.

En una Normal, esos serían la media (μ) y la desviación típica (σ).

En la Cauchy, como no tiene media ni varianza, ¿qué son?

- **a es la Mediana (y la Moda):** Es el punto donde está el pico de la campana. El 50% de los datos caen a su izquierda y el 50% a su derecha. Pero no es el centro de gravedad (media).
- **b es el Parámetro de Escala:** Te dice lo “gorda” que es la campana, pero no es la desviación típica. De hecho, b mide la mitad del Rango Intercuartílico.

La propiedad tipificadora $Y = (X - a)/b \sim Ca(0, 1)$ funciona igual que en la Normal para “centrar y escalar” la campana, pero conceptualmente a y b no son momentos.

4. La Física del Arcotangente (La Función de Distribución)

La función de distribución acumulada es: $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctg\left(\frac{x-a}{b}\right)$.

¿Por qué narices aparece un arco-tangente en la probabilidad? Aquí viene la intuición más hermosa de todas, conocida como **La Paradoja de la Linterna**:

Imagina que estás vendado en el centro de una habitación infinita. Tienes una linterna en la mano que gira locamente en todas direcciones. Te paras al azar y enciendes la linterna hacia una pared recta infinita que está a b metros de ti.

El ángulo en el que apuntas la linterna es completamente Uniforme (entre -90° y $+90^\circ$).

¿Dónde cae el punto de luz en la pared (la coordenada x)?

¡Por trigonometría, la posición del punto de luz es $x = b \cdot \tan(\theta)$!

Si el ángulo θ es Uniforme, la posición x sigue exactamente una distribución de Cauchy.

Si apuntas casi paralelo a la pared (ángulo cercano a 90°), el punto de luz se va a millones de kilómetros (ahí tienes tu Cisne Negro, el evento extremo).

Como la relación es la tangente, para recuperar la probabilidad acumulada tienes que hacer la inversa: **¡el arcotangente!** Por eso la $F(x)$ tiene un $\arctg$.

1. Tu “falsa Gaussiana” y el Secreto de Taylor

Sabemos que $e^u \approx 1 + u$.

Por tanto, $e^{x^2} \approx 1 + x^2$.

Si invertimos ambos lados:

$$e^{-x^2} = \frac{1}{e^{x^2}} \approx \frac{1}{1 + x^2}$$

¡BOOM! Cerca del origen (donde x es pequeño), la Gaussiana y la Cauchy son exactamente la misma curva. Tu cerebro buscó la función racional (fracción) más simple y elegante que formara una campana simétrica. Acertaste de pleno. La única diferencia entre ambas ocurre cuando x se hace grande (las colas), donde la exponencial se desploma y la fracción sobrevive.

2. La Historia Real: Un arma de destrucción matemática

¿Fue heurística por el arcotangente? Sí y no.

La curva $\frac{1}{1+x^2}$ ya existía en geometría pura mucho antes que en probabilidad. La estudió Fermat en 1630 y la hizo famosa la matemática italiana Maria Gaetana Agnesi en 1748. En geometría se la conoce con el poético (y mal traducido) nombre de *La Bruja de Agnesi*.

Pero su entrada en la probabilidad no fue porque “el arcotangente pareciera una CDF”. Fue un ataque premeditado.

En 1824 (mucho antes de que Cauchy la popularizara en 1853), el matemático Siméon Denis Poisson estaba cabreado. Laplace acababa de formular el sagrado Teorema Central del Límite (TCL), afirmando que “todo tiende a la campana de Gauss”.

Poisson quería demostrar que Laplace se equivocaba, que el TCL tenía límites. Necesitaba encontrar un “monstruo” matemático que:

- Tuviera forma de campana (para engañar a primera vista).
- Su área fuera exactamente 1 (para ser una probabilidad válida).
- Su varianza fuera infinita (para romper las condiciones del TCL).

Poisson rebuscó en el baúl de la geometría, sacó a la Bruja de Agnesi, la normalizó dividiendo por π (cuya primitiva es, en efecto, el arcotangente) y demostró que si sumas variables con esta distribución, jamás convergen a la Normal. Fue creada literalmente como un “anti-teorema”.

3. El Jugo de la Cauchy: Tres Intuiciones Abismales

Ya que hemos abierto la caja de Pandora, te voy a dar tres propiedades de la Cauchy que te van a volar la cabeza.

A. El Cociente de dos Normales (El origen del caos)

Imagina que tienes dos variables X e Y , ambas Normales estándar perfectas $N(0, 1)$. El orden puro.

¿Qué pasa si las divides? $Z = X/Y$.

Al dividir, a veces la variable Y tomará un valor muy, muy cercano a cero (porque la Normal tiene mucha masa en el centro). Dividir X entre un número minúsculo catapulta el resultado al infinito.

El cociente de dos Gaussianas independientes es exactamente una distribución de Cauchy. **Intuición:** El caos de la Cauchy nace del simple acto de dividir por cero en un mundo Normal.

B. La Inmunidad a la Ley de los Grandes Números (El Agujero Negro)

Si tú mides la altura de 10 personas y haces la media, tienes un dato fiable. Si mides a 1.000, tienes un dato hiperpreciso. La varianza de la media muestral se reduce a medida que n crece (se divide por n).

Si tus datos siguen una distribución de Cauchy y haces la media de 1.000 observaciones... esa media sigue siendo una distribución de Cauchy idéntica a la original.

Intuición: ¡La media no aprende! Tener 1 millón de muestras de una Cauchy te da exactamente la misma incertidumbre que tener 1 sola muestra. El sumatorio de los datos crece al mismo ritmo monstruoso que n . Es la distribución “fractal” por excelencia; su promedio es un clon de sí misma.

C. La Ruleta Rusa Geométrica (Proyección Estereográfica)

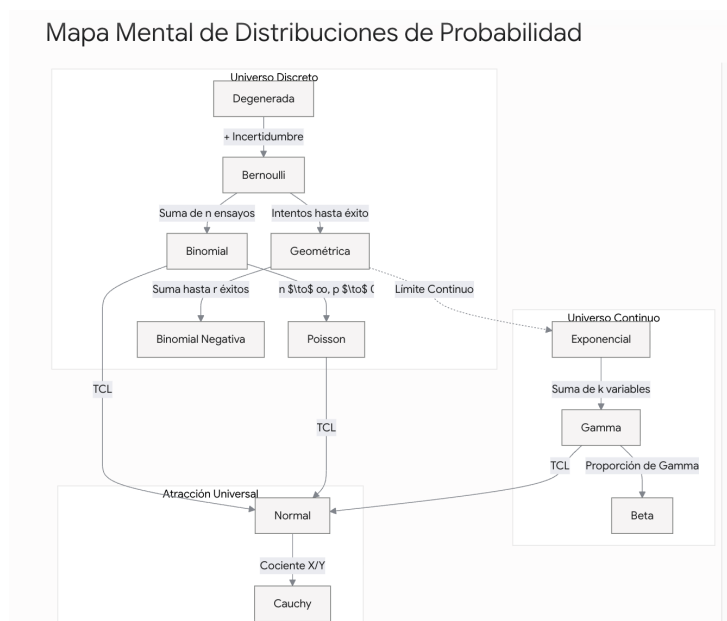
Imagina un círculo perfecto apoyado sobre el eje X (como una rueda sobre el suelo). Pones un láser en el punto más alto de la rueda (el polo norte).

Ahora, eliges un punto totalmente al azar en la circunferencia (probabilidad Uniforme) y disparas el láser desde el polo norte, pasando por ese punto aleatorio, hasta que el rayo choque contra el suelo (el eje X).

¿Cuál es la distribución estadística de los puntos de impacto en el suelo? Es una distribución de Cauchy.

Intuición: Si el punto aleatorio cae cerca del polo sur, el láser da casi en el centro. Pero si el punto aleatorio cae cerca del ecuador, el láser sale disparado casi paralelo al suelo y el impacto se va a cientos de kilómetros. La Cauchy es simplemente la “sombra” unidimensional de una probabilidad uniforme en un círculo bidimensional.

1.11 Esquema general y deducción ultra rápida de las distribuciones



1. Binomial Negativa (el número de ensayos para r éxitos)

Olvídate de fórmulas complejas. Piensa en la cronología del evento. Quieres saber la probabilidad de que necesites exactamente k ensayos para conseguir r éxitos.

Condición innegociable: para que el juego termine exactamente en el paso k , el último intento (k) tiene que ser un éxito obligatoriamente. La probabilidad de ese último éxito es p .

Lo que pasó antes: si el último fue éxito y en total hay r éxitos, significa que en los $k - 1$ ensayos anteriores tuviste que conseguir exactamente $r - 1$ éxitos. ¡Y esto es una Binomial normal y corriente!

Multiplica ambas cosas:

$$\mathbb{P}(X = k) = \underbrace{\binom{k-1}{r-1} p^{r-1} (1-p)^{(k-1)-(r-1)}}_{\text{Binomial en los } k-1 \text{ pasos previos}} \times \underbrace{p}_{\text{Éxito final}}$$

Agrupas las p y tienes la fórmula $\mathbb{P}(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$. Demostrado en 10 segundos.

1. La diferencia de convención

Mi deducción (total de ensayos): yo definí la variable como el número **total** de tiradas necesarias para conseguir r éxitos. En mi fórmula, si quieres 3 éxitos, lo mínimo que puedes tirar son 3 veces. El soporte empieza en $r, r+1, r+2, \dots$

Tus apuntes (total de fracasos): tus apuntes definen la variable X como el número **total** de fracasos antes de conseguir los r éxitos. Si tienes mucha suerte, puedes tener 0 fracasos. El soporte empieza en $0, 1, 2, \dots$

2. La transformación matemática (hazlo al instante)

Sea n mi variable (total de tiradas) y sea k la variable de tus apuntes (total de fracasos). Es una obviedad física que:

$$\text{Total de tiradas} = \text{Fracasos} + \text{Éxitos} \quad \implies \quad n = k + r$$

Coge la fórmula que te deduje antes para n tiradas:

$$\mathbb{P} = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

Ahora, sustituye estrictamente la n por $k+r$:

$$\mathbb{P} = \binom{(k+r)-1}{r-1} p^r (1-p)^{(k+r)-r}$$

Simplificamos los exponentes del fracaso $(1-p)$, que tus apuntes llaman q :

$$\mathbb{P} = \binom{k+r-1}{r-1} q^k p^r$$

3. El truco final de combinatoria

A ti te queda el número combinatorio $\binom{k+r-1}{r-1}$ y en tus apuntes pone $\binom{k+r-1}{k}$. ¿Son distintos? No. Por la propiedad de simetría de los números combinatorios $\binom{N}{A} = \binom{N}{N-A}$:

$$\binom{k+r-1}{r-1} = \binom{k+r-1}{(k+r-1)-(r-1)} = \binom{k+r-1}{k}$$

¡Boom! Es exactamente la misma ecuación.

2. La Gamma (el tiempo hasta el r -ésimo evento)

Tú mismo lo dijiste: la Gamma es la versión continua de la Binomial Negativa. Es el tiempo T que tardas en ver r destellos radiactivos (eventos de Poisson). Hagamos la misma cronología que antes en un intervalo continuo de tiempo hasta t :

Condición innegociable: para que el r -ésimo destello ocurra exactamente en el instante t (en una franja infinitesimal dt), la probabilidad es la tasa por el diferencial: λdt .

Lo que pasó antes: en todo el tiempo anterior t , tuvieron que ocurrir exactamente $r - 1$ destellos. ¿Quién da el número de eventos en un tiempo t ? Poisson. Probabilidad de $r - 1$ eventos: $\frac{(\lambda t)^{r-1} e^{-\lambda t}}{(r-1)!}$.

Multiplica ambas cosas (densidad $\times dt$):

$$f(t) dt = \underbrace{\frac{(\lambda t)^{r-1} e^{-\lambda t}}{(r-1)!}}_{\text{Poisson antes de } t} \times \underbrace{\lambda dt}_{\text{Destello en } t}$$

Junta las λ : $f(t) = \frac{\lambda^r t^{r-1} e^{-\lambda t}}{(r-1)!}$. Cambia el factorial $(r - 1)!$ por la función continua $\Gamma(r)$ y ya tienes la Gamma. Demostrado geométricamente.

3. La Beta (la proporción o la “probabilidad de la probabilidad”)

La Beta es una distribución sobre el intervalo $[0, 1]$. Mide proporciones (ej. “cuál es la probabilidad de que una moneda esté trucada”). Imagina que tiras una moneda y obtienes $\alpha - 1$ caras y $\beta - 1$ cruces. ¿Cuál es la ecuación de la probabilidad de que el parámetro de éxito sea x ? Pues x elevado al número de caras, por $(1 - x)$ elevado al número de cruces:

$$\text{Núcleo} = x^{\alpha-1} (1-x)^{\beta-1}$$

Como es una probabilidad, el área bajo la curva tiene que sumar 1. Así que divides por la integral de todo eso (que por definición se llama Función Beta $B(\alpha, \beta)$):

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

Directo de la lógica de Bayes.

4. La Normal (la campana del error)

Es la más difícil de derivar algebraicamente desde cero, pero la más fácil de intuir. Imagina que disparas a una diana; el error se distribuye en el espacio.

- El error en el eje X es independiente del error en el eje Y : $f(x, y) = f(x)f(y)$.
- La probabilidad solo depende de la distancia al centro $r^2 = x^2 + y^2$, es decir, la forma es circular/isotrópica: $f(x)f(y) = g(x^2 + y^2)$.
- La única función matemática en todo el universo que cumple que $f(x)f(y) = g(x^2 + y^2)$ es la función exponencial (porque $e^A e^B = e^{A+B}$).

Por tanto, la densidad tiene que ser proporcional a e^{-x^2} . Normalizas para que el área sea 1, le metes la varianza σ para escalar, y te escupe la Normal.

1.11.1 Deducción rápida de Beta a partir de Gamma

El problema: tienes el núcleo $x^{\alpha-1} (1-x)^{\beta-1}$ y necesitas que el área bajo él valga 1. Por pura definición, la Función Beta es esa área integral:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

La relación curiosa (la deducción legendaria):

¿Cómo se relaciona esto con Γ ? Piensa en las combinatorias. El coeficiente binomial es $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Si Γ es el factorial, la función Beta (que mide proporciones y combinaciones continuas) debería ser un cociente de funciones Gamma. Vamos a demostrarlo en tres líneas.

Multiplica dos funciones Gamma (usa variables u y v):

$$\Gamma(\alpha)\Gamma(\beta) = \left(\int_0^\infty u^{\alpha-1} e^{-u} du \right) \left(\int_0^\infty v^{\beta-1} e^{-v} dv \right)$$

Como son independientes, las juntamos en una integral doble:

$$= \int_0^\infty \int_0^\infty u^{\alpha-1} v^{\beta-1} e^{-(u+v)} du dv$$

El truco genial de la proporción: si u y v son cantidades absolutas, definamos una cantidad total $t = u + v$ y una proporción $x = \frac{u}{u+v}$. De aquí despejamos $u = tx$ y $v = t(1-x)$. El Jacobiano de este cambio de variable es t , por lo que $du dv = t dx dt$. Sustituimos esto en la integral doble:

$$= \int_0^\infty \int_0^1 (tx)^{\alpha-1} (t(1-x))^{\beta-1} e^{-t} (t) dx dt$$

Agrupamos todas las t y todas las x , y separamos las integrales:

$$= \left(\int_0^\infty t^{\alpha+\beta-1} e^{-t} dt \right) \times \left(\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx \right)$$

Mira fijamente esos dos paréntesis: ¡son las definiciones exactas que forjamos antes!

- El primer paréntesis es estrictamente $\Gamma(\alpha + \beta)$.
- El segundo paréntesis es estrictamente $B(\alpha, \beta)$.

Por tanto:

$$\Gamma(\alpha)\Gamma(\beta) = \Gamma(\alpha + \beta) \cdot B(\alpha, \beta)$$

Despejando:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad \blacksquare$$

Intuición brutal: el universo total de posibilidades (el producto de los factoriales independientes $\Gamma(\alpha)\Gamma(\beta)$) se divide topológicamente de forma perfecta en dos entes separados: la escala total o “tiempo total” ($\Gamma(\alpha + \beta)$) multiplicada por la fracción o proporción del sistema repartida en $[0, 1]$ (la Función Beta).

1.11.2 Deducción rápida de la FGM de Gamma

Línea 1: planteamiento y agrupación de exponentes.

Metemos la exponencial en la integral de la densidad y agrupamos las x en el exponente:

$$M_X(t) = \int_0^\infty e^{tx} \left(\frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} \right) dx = \frac{\lambda^p}{\Gamma(p)} \int_0^\infty x^{p-1} e^{-(\lambda-t)x} dx$$

Línea 2: el truco analítico (reconocimiento del núcleo).

Mira fijamente la integral que ha quedado: $\int_0^\infty x^{p-1} e^{-(\lambda-t)x} dx$. ¡Ese es exactamente el núcleo de una nueva distribución Gamma! Su parámetro de forma sigue siendo p , pero su nuevo parámetro de tasa ha mutado a $\lambda^* = (\lambda - t)$.

¿Cuánto vale la integral del núcleo de una Gamma de parámetros p y λ^* ? Vale exactamente su constante de normalización invertida (el área total que dedujimos antes): $\frac{\Gamma(p)}{(\lambda^*)^p}$.

Condición innegociable: para que esto sea una Gamma válida y el área no explote al infinito, el nuevo parámetro de tasa tiene que ser positivo. Es decir, $\lambda - t > 0 \implies t < \lambda$. ¡Acabas de descubrir el dominio de convergencia de la FGM!

Línea 3: sustitución y colapso algebraico.

Sustituimos la integral entera por ese valor:

$$M_X(t) = \frac{\lambda^p}{\Gamma(p)} \times \left[\frac{\Gamma(p)}{(\lambda - t)^p} \right]$$

Las $\Gamma(p)$ se aniquilan entre sí. Agrupamos el exponente p común y... ¡Boom! FGM de la Gamma obtenida:

$$M_X(t) = \left(\frac{\lambda}{\lambda - t} \right)^p, \quad \text{para } t < \lambda \quad \blacksquare$$

Distribución multinomial

2.1 Intuición

«En un caso binomial, como da la “casualidad” de que conocido p, q es obtenible, la “distribución inversa” (ahora digo a qué me refiero) se obtiene determinísticamente. Queriendo decir que, si yo en n tiradas de un dado, saber la probabilidad de haber obtenido x aciertos implica inmediatamente que he obtenido $n - x$ fallos, y viceversa; i.e., si yo definiera un vector aleatorio (pues entiendo que las distribuciones “multi” son para vectores en lugar de variables aleatorias, ¿es esa la diferencia?) (X_1, X_2) , la función de probabilidad de distribución conjunta no posee más información que la de X_1 solo. Eso naturalmente se rompe cuando tengo tres “paquetes de opciones”, tres “clases” en las que descomponer mi evento.

Aquí tengo la confusión de saber exactamente cómo debe verse, porque yo tengo la idea casi temporal de primero tirar una moneda, luego otra, luego otra, y voy construyendo la binomial. Pero ahora quizás tenga más sentido decir: “mira, tengo n monedas, las tiro a la vez y distribuyo en las dos clases”. Y en este caso, pues por ejemplo sería la forma canónica de verlo, quizás según me imagino, pensar en un dado de j caras con p_j de probabilidad para cada una de las caras (arbitrarias pero sumando 1, lo cual me da algo que en cierto sentido, al ser convexo, debería tener que ver con análisis convexo, ¿no?, aunque esto puede ser una inventada. Ya sé que hay por ahí una visión del simplex pero no sé si me merece meterme ahí ahora).

Y entonces la multinomial de j clases y n tiradas (que no sé cuál es el nombre formal) sería imaginar mi vector aleatorio de 6 dimensiones y asociar a cada distribución posible (v.gr., 1, 4, 2, 3, 0, 2) su probabilidad asociada, ¿me equivoco? Y esa probabilidad pues es trivial: es el producto de las probabilidades por el factor este de $n!$ dividido entre los factoriales de cada uno de los elementos que hay en las clases, ¿no?»

1. La Dimensionalidad y la Redundancia: De la Variable al Vector

Tienes toda la razón. La diferencia fundamental entre la distribución binomial y la multinomial es el salto de una variable aleatoria a un vector aleatorio. En la binomial, tu espacio de resultados se divide en dos clases (éxito y fracaso). Como bien intuyes, si tienes n ensayos y conoces que hay X_1 éxitos, automáticamente tienes $n - X_1$ fracasos. El grado de libertad es 1. Por eso, la distribución binomial clásica se modela con una simple variable aleatoria unidimensional en $\mathbb{R}$.

Cuando saltas a k clases (tu dado de j caras), esta “casualidad” se rompe parcialmente. Necesitas un vector aleatorio k -dimensional $\underline{X} = (X_1, \dots, X_k)$ donde cada X_i es el número de veces que ocurre el suceso A_i en las n repeticiones. Sin embargo, la redundancia no desaparece del todo, solo se traslada. Al igual que en la binomial, la última componente está fijada por las demás porque $\sum_{i=1}^k X_i = n$. Es decir, un vector multinomial de k componentes tiene, topológicamente hablando, $k - 1$ grados de libertad. Esto nos lleva directamente a tu genial intuición sobre la convexidad.

2. La Geometría del Espacio Muestral: El Símplex

Tu mención al análisis convexo y al símplex no es una “inventada”; es la verdadera geometría subyacente de la teoría de la probabilidad. Entremos de lleno ahí. Imagina un espacio tridimensional (un dado de 3 caras, $k = 3$). Las probabilidades de cada cara (p_1, p_2, p_3) deben ser no negativas y sumar 1. Si graficas esto en $\mathbb{R}^3$, la ecuación $p_1 + p_2 + p_3 = 1$ no llena un volumen tridimensional; define un hiperplano afín bidimensional que corta los ejes en $(1, 0, 0)$, $(0, 1, 0)$ y $(0, 0, 1)$. El triángulo equilátero que se forma al unir estos tres puntos es el símplex estándar Δ^2 . Toda la teoría de probabilidad discreta vive en este símplex. Es un objeto fundamental en topología algebraica.

Ahora, pensemos en los resultados, el vector de tu distribución posible (x_1, x_2, x_3) . Estos también deben ser enteros no negativos y sumar n . Geométricamente, el soporte de la distribución multinomial es una red cristalina (lattice) discreta que vive sobre un símplex dilatado por un factor de n . Cuando evalúas la función de probabilidad conjunta, no estás calculando probabilidades a lo largo de un eje continuo (como harías en el cálculo multivariable clásico), sino evaluando “pesos” en los nodos de esta red discreta sobre el hiperplano. Debido a que el vector está confinado a este hiperplano $\sum X_i = n$, la matriz de varianzas y covarianzas de la distribución multinomial es singular (su determinante es 0); no tiene “volumen” real en $\mathbb{R}^k$.

3. La Analogía Física: El Tiempo, el Espacio y la Termodinámica

Mencionas la tensión entre la visión “temporal” (tiro monedas una a una) y la visión “espacial/simultánea” (tiro n monedas a la vez). Matemáticamente, son equivalentes porque el experimento asume que las repeticiones son independientes y las probabilidades permanecen constantes. Sin embargo, tu visión simultánea (“tengo n monedas, las tiro a la vez y distribuyo en clases”) es profundamente física.

De hecho, es la base de la Mecánica Estadística. Imagina que en lugar de tirar n dados, tienes un gas con n partículas (átomos) y k niveles de energía posibles (las “caras” del dado o “clases”). Si las partículas no interactúan y se distribuyen al azar, la forma en que se reparten en los niveles de energía sigue exactamente una distribución multinomial.

4. La Estructura de la Probabilidad: El Coeficiente Multinomial

Tu última frase resume la fórmula perfectamente. La probabilidad de obtener un vector

específico $(x_1, \dots, x_k)$ es el producto de las probabilidades de cada evento $(p_1^{x_1} p_2^{x_2} \dots p_k^{x_k})$ multiplicado por el factor combinatorio $\frac{n!}{x_1! \dots x_k!}$.

¿Por qué existe este factor? Veámoslo con teoría de grupos. El producto de las probabilidades aisladas (e.g., $p_1^{x_1} \dots p_k^{x_k}$) es la probabilidad de que ocurra una secuencia específica e inmutable en el tiempo (e.g., primero todos los 1s, luego los 2s, etc.). Pero a tu vector aleatorio multinomial no le importa el orden en el tiempo (el sumatorio espacial colapsa el orden temporal).

Ese factor de factoriales, matemáticamente conocido como el coeficiente multinomial $\binom{n}{x_1, \dots, x_k}$, mide el tamaño de la órbita bajo la acción del grupo simétrico S_n . Es decir, calcula exactamente de cuántas formas diferentes (permutaciones espaciales) puedes entrelazar esos resultados de modo que el conteo final de clases siga siendo idéntico. Se trata del número de microestados que dan lugar al mismo macroestado. En física, el logaritmo de este coeficiente multinomial es literalmente lo que Ludwig Boltzmann definió como Entropía.

«No veo exactamente en qué me ayuda ni teórica ni prácticamente el hecho de saber que, por un lado, vive en una red cristalina del simplex multiplicado por n (el soporte); me parece completamente irrelevante.

Y por otro lado: “Imagina un espacio tridimensional (un dado de 3 caras, $k = 3$). Las probabilidades de cada cara (p_1, p_2, p_3) deben ser no negativas y sumar 1. Si graficas esto en $\mathbb{R}^3$, la ecuación $p_1 + p_2 + p_3 = 1$ no llena un volumen tridimensional; define un hiperplano afín bidimensional que corta los ejes en $(1, 0, 0)$, $(0, 1, 0)$ y $(0, 0, 1)$. El triángulo equilátero que se forma al unir estos tres puntos es el simplex estándar Δ^2 . Toda la teoría de probabilidad discreta vive en este simplex. Es un objeto fundamental en topología algebraica.” Esto también me parece inútil, pues para una distribución concreta de todos modos p_1 , p_2 y p_3 están definidos de forma concreta. ¿Qué más me da que si los hiciera libres habitaran ese simplex?

Dices por otro lado lo de la covarianza de det. cero. Vamos a ver. Para mí la matriz de covarianzas dado un vector aleatorio $X_1, X_2, X_3, \dots, X_n$, es simplemente la matriz de Gram de proyectar otra variable aleatoria Y sobre el subespacio generado por $X_1, X_2, \dots, X_n$, esa es la intuición que yo tengo, otra no (no sé si es la que hay que tener).

Pero entonces no sé qué narices tiene ahí que decirme que el determinante sea cero... Naturalmente en ese caso esas funciones no forman una base... de ese espacio, pues es un subespacio de dimensión $N - 1$ (y tenemos N variables aleatorias). También hay una cierta visión del tensor métrico para la matriz de covarianzas, y eso me diría que es degenerado... »

Si $p_1, \dots, p_k$ están fijos y dados por Dios, hablar del simplex no te aporta nada práctico para calcular una probabilidad.

Pero en la matemática real y aplicada (Estadística, Teoría de la Información y Machine Learning), los parámetros p_i nunca son conocidos. El problema central no es calcular la probabilidad dado un p fijo, sino encontrar el p óptimo dadas unas observaciones. Ahí es donde entra el análisis convexo con una violencia brutal.

$E = \mathbb{R}^k$. Se define el simplex estándar de probabilidad de dimensión $k - 1$ como el conjunto convexo y compacto:

$$\Delta^{k-1} = \{ \underline{p} \in \mathbb{R}^k : p_i \geq 0 \text{ para } i = 1, \dots, k, \text{ y } \sum_{i=1}^k p_i = 1 \}$$

Nota Filosófica/Práctica: Como bien dices, para una distribución concreta, $\underline{p}$ es un punto fijo. ¿Para qué hacerlos libres? Porque en inferencia estadística (por ejemplo, estimación de máxima verosimilitud), tú diseñas una función de pérdida (la entropía cruzada) y debes minimizarla. Optimizar funciones sobre variedades requiere conocer la geometría del dominio. Tu dominio de optimización no es $\mathbb{R}^k$, es Δ^{k-1} .

Si en estadística el simplex es el terreno de caza donde buscas el parámetro oculto, en probabilidad el simplex es la jaula topológica que define el comportamiento asintótico de tu sistema. Imagínate la Ley de los Grandes Números y el Teorema Central del Límite. Cuando tú repites el experimento multinomial n veces y divides por n , obtienes el vector de frecuencias relativas: $\frac{\underline{X}}{n} = (\frac{X_1}{n}, \dots, \frac{X_k}{n})$. A medida que $n \rightarrow \infty$, esta nube microscópica de probabilidades colapsa. ¿Hacia dónde colapsa? Hacia un único punto exacto: el vector $\underline{p}$.

Pero la probabilidad no estudia solo el límite, estudia las fluctuaciones alrededor de ese límite (el Teorema Central del Límite). Las fluctuaciones de una multinomial no pueden expandirse en una esfera perfecta de $\mathbb{R}^k$. Están forzadas a vibrar microscópicamente exclusivamente a lo largo de la superficie del simplex. La geometría del simplex define los “modos normales de vibración” de tu sistema estocástico. Si no entiendes que la probabilidad vive incrustada en ese hiperplano, no podrás entender por qué la distribución Normal Multivariante asintótica tiene la forma aplastada que tiene.

Imposibilidad de la Función de Densidad (Singularidad):

En tu propio temario, sección 2.1.3, te hablan de Vectores aleatorios singulares. Un vector con $\det(\Sigma) = 0$ (en su versión continua) no tiene función de densidad con respecto a la medida de Lebesgue en $\mathbb{R}^k$. Es físicamente imposible escribir su fórmula $f(x_1, \dots, x_k)$. Al igual que no puedes calcular la “densidad de masa tridimensional” de una hoja de papel ideal (sería infinito en el papel y cero fuera), no puedes integrar probabilidades usando volumen en k dimensiones. Tienes que usar teoría de la medida para reducir la dimensionalidad o usar densidades generalizadas (deltas de Dirac).

El Colapso de la Normal Multivariante (Cálculo Práctico):

Si intentas aplicar el Teorema Central del Límite a la multinomial para aproximarla a una Normal Multivariante en $\mathbb{R}^k$, vas a escribir la fórmula de la Normal:

$$f(\underline{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \Sigma^{-1}(\underline{x} - \underline{\mu})\right)$$

¡Boom! División por cero en el denominador por el $\det(\Sigma)$, y error matemático al intentar calcular Σ^{-1} en el exponente porque una matriz singular no tiene inversa. Saber que el determinante es nulo te avisa de antemano de que tu modelo matemático estallará si no eliminas la variable redundante (proyectando el sistema de k a $k - 1$ dimensiones).

Filosofía de la Información (Entropía Cero en un eje):

Que el determinante sea nulo te dice que tu sistema tiene una “falsa libertad”. Crees que estás estudiando k dimensiones de incertidumbre, pero el universo te está obligando a moverte en $k - 1$. Si diseñas una simulación computacional de este vector, saber esto te permite ahorrar un $O(k^3)$ en complejidad computacional al invertir matrices, descartando la dimensión vacía.

2.2 Vector aleatorio multinomial

Supongamos un experimento con $k > 1$ posibles resultados, $A_1, A_2, \dots, A_k$, que son sucesos de un cierto espacio de probabilidad que son mutuamente excluyentes ($A_i \cap A_j = \emptyset, i \neq j$) y exhaustivos ($\bigcup_{i=1}^k A_i = \Omega$). Este experimento se repite n veces, de tal forma que las probabilidades $p_i = \mathbb{P}(A_i), i = 1, \dots, k$ permanecen constantes y que los resultados de las diferentes repeticiones sean independientes.

2.2.1 Definición

En las condiciones anteriores, el vector aleatorio k dimensional $\underline{X} = (X_1, \dots, X_k)'$ con:

$$X_i = \text{número de veces que ocurre } A_i \text{ en las } n \text{ repeticiones, } i = 1, \dots, k$$

es un vector aleatorio multinomial y se denota $\underline{X} \sim \mathcal{M}(n; p_1, \dots, p_k)$.

Obsérvese que $\sum_{i=1}^k p_i = 1$.

2.2.2 Teorema (función de probabilidad) *

La función de probabilidad de $\underline{X} = (X_1, X_2, \dots, X_k)' \sim \mathcal{M}(n; p_1, \dots, p_k)$ es:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k},$$

para $x_i \geq 0, i = 1, \dots, k$ enteros tales que $\sum_{i=1}^k x_i = n$.

Demostración. El suceso $\{X_1 = x_1, \dots, X_k = x_k\}$ significa que el suceso A_1 ocurre x_1 veces, A_2 ocurre x_2 veces, $\dots$, A_k ocurre x_k veces, entre las n repeticiones independientes del experimento. Obviamente, $x_i \geq 0$ debe ser entero y $\sum_{i=1}^k x_i = n$. Entonces debido a la independencia:

$$\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = C p_1^{x_1} p_2^{x_2} \dots p_k^{x_k},$$

donde C es el número de formas diferentes en las que aparecen x_i repeticiones del suceso $A_i, i = 1, \dots, k$, entre las n repeticiones del experimento. (C obviamente depende de $x_1, \dots, x_k$).

Para calcular C , supongamos que tenemos n posiciones en las que colocaremos los símbolos A_i x_i veces, $i = 1, \dots, k$. C es precisamente el número total de ordenaciones de este tipo que puedan hacerse. Para el conteo, procedemos secuencialmente:

- Colocar x_1 símbolos A_1 en las n posiciones: esto puede hacerse de $\binom{n}{x_1}$ formas.
- Colocar x_2 símbolos A_2 en las $n - x_1$ posiciones libres: esto puede hacerse de $\binom{n-x_1}{x_2}$ formas.
- Continuar secuencialmente de esta manera hasta que se coloquen todos los símbolos en las n posiciones,

entonces:

$$C = \binom{n}{x_1} \binom{n-x_1}{x_2} \dots \binom{n-x_1-\dots-x_{k-1}}{x_k} = \frac{n!}{x_1! x_2! \dots x_k!}. \quad \blacksquare$$

2.2.3 Componente redundante y binomio de Newton generalizado

En realidad una de las componentes del vector aleatorio multinomial es redundante, pues las componentes satisfacen la relación $\sum_{i=1}^k X_i = n$. Por éste motivo, algunos autores suprimen una componente, por ejemplo la última y consideran el vector de $k - 1$ componentes

$(X_1, \dots, X_{k-1})$ en vez del vector con k componentes tal como en la Definición anterior. La relación entre ambas notaciones es clara:

$$\mathbb{P}(X_1 = x_1, \dots, X_{k-1} = x_{k-1}) = \mathbb{P}\left(X_1 = x_1, \dots, X_{k-1} = x_{k-1}, X_k = n - \sum_{i=1}^{k-1} x_i\right).$$

La siguiente fórmula es una generalización del binomio de Newton:

$$(a_1 + \dots + a_k)^n = \sum_{\substack{x_1, \dots, x_k \\ x_i \geq 0 \\ \sum_i x_i = n}} \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k a_i^{x_i},$$

donde:

$$\binom{n}{x_1, \dots, x_k} = \frac{n!}{x_1! \dots x_k!}, \quad \text{siempre que } \sum_{i=1}^k x_i = n.$$

La generalización de la fórmula del binomio de Newton es útil para probar que la expresión obtenida en el Teorema anterior es una función de probabilidad. Obsérvese que si $\underline{X}$ es un vector aleatorio multinomial:

$$\text{Sop}(\underline{X}) = \left\{ (x_1, \dots, x_k) : x_i \geq 0, \text{ enteros}, i = 1, \dots, k, \sum_{i=1}^k x_i = n \right\},$$

entonces:

$$\begin{aligned} \sum_{(x_1, \dots, x_k) \in \text{Sop}(\underline{X})} \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) &= \sum_{\substack{x_1, \dots, x_k \\ x_i \geq 0 \\ \sum_i x_i = n}} \binom{n}{x_1, \dots, x_k} p_1^{x_1} \dots p_k^{x_k} \\ &= (p_1 + \dots + p_k)^n = 1. \end{aligned}$$

2.3 El Teorema Multinomial

Existen dos direcciones ortogonales en las que se puede generalizar el Binomio clásico $(a + b)^n$:

Generalización en el dominio del exponente (El verdadero Binomio Generalizado de Newton): Isaac Newton, en 1665, se preguntó qué pasaba si n no era un número natural, sino un número real cualquiera (negativo, fraccionario, irracional). Para ello, extendió los coeficientes binomiales usando factoriales descendentes (lo que hoy entendemos mediante la función Gamma) y transformó el polinomio finito en una serie infinita (la serie de Taylor de $(1 + x)^a$). Tú te referías a este, y es el nombre históricamente correcto.

Generalización en la dimensión de la base (El Teorema Multinomial): ¿Qué pasa si en lugar de un binomio $(a + b)$ tenemos un polinomio de k términos $(a_1 + \dots + a_k)$ elevado a un entero n ? Esta fórmula fue sistematizada por Gottfried Leibniz. El nombre formal, universal y riguroso de la fórmula de tu imagen es Teorema Multinomial.

Lamarlo “generalización del binomio de Newton” no es matemáticamente falso (pues generaliza la base), pero es pedagógicamente torpe porque usurpa el nombre de un teorema analítico infinitamente más profundo (el de los exponentes reales).

2.3.1 Teorema (Teorema Multinomial)

Para cualquier k -tupla de números reales (o elementos de un anillo conmutativo) $(a_1, \dots, a_k)$ y cualquier entero $n \geq 1$, se verifica que:

$$(a_1 + \dots + a_k)^n = \sum_{(x_1, \dots, x_k) \in \Omega_{n,k}} \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k a_i^{x_i}$$

Demostración (Vía Mecánica Combinatoria). Consideremos el producto expansivo de n factores idénticos:

$$(a_1 + a_2 + \dots + a_k) \cdot (a_1 + a_2 + \dots + a_k) \cdots (n \text{ veces})$$

Por la propiedad distributiva, al expandir este producto masivo, cada término individual se forma eligiendo exactamente una variable a_i de cada uno de los n paréntesis, y luego multiplicándolas juntas.

Supongamos que en una elección arbitraria hemos escogido la variable a_1 exactamente x_1 veces, la variable a_2 exactamente x_2 veces, y así sucesivamente hasta a_k . Evidentemente, la suma de las elecciones debe ser el total de paréntesis: $\sum_{i=1}^k x_i = n$. El término algebraico resultante de esta elección será de la forma:

$$a_1^{x_1} a_2^{x_2} \dots a_k^{x_k}$$

Ahora, la pregunta geométrica y combinatoria es: ¿Cuántas “rutas” diferentes a través de los n paréntesis generan exactamente este mismo término algebraico $a_1^{x_1} \dots a_k^{x_k}$?

Esto equivale a preguntar: De los n paréntesis disponibles (que podemos imaginar como n “ranuras temporales” o “cajas” ordenadas), ¿de cuántas formas puedo elegir x_1 ranuras para meter la letra a_1 , luego de las $n - x_1$ ranuras restantes elegir x_2 para la letra a_2 , y así sucesivamente?

Calculamos este producto encadenado de combinaciones clásicas:

$$\binom{n}{x_1} \cdot \binom{n-x_1}{x_2} \cdot \binom{n-x_1-x_2}{x_3} \cdots \binom{x_k}{x_k}$$

Si expandimos los factoriales de esta cadena de productos, ocurre una cancelación telescópica masiva:

$$\frac{n!}{x_1!(n-x_1)!} \cdot \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} \cdots = \frac{n!}{x_1!x_2! \dots x_k!}$$

Esta cantidad es exactamente el coeficiente multinomial $\binom{n}{x_1, \dots, x_k}$. Como esto es cierto para toda partición posible $(x_1, \dots, x_k)$ que sume n , al agrupar términos semejantes obtenemos la fórmula del teorema. ■

2.4 Relación con la distribución binomial

2.4.1 Caso $k = 2$

En el caso de que $k = 2$, se obtiene la distribución binomial.

Demostración. En efecto, consideremos un experimento con dos posibles resultados: A_1 y A_2 , excluyentes, $A_1 \cap A_2 = \emptyset$, y exhaustivos $A_1 \cup A_2 = \Omega$, con probabilidades respectivas $p_i = \mathbb{P}(A_i)$, $i = 1, 2$. Obviamente $p_1 + p_2 = 1$. Si el experimento se repite n veces independientemente, el correspondiente vector multinomial es $\underline{X} = (X_1, X_2)$ con:

$$X_i = \text{número de veces que aparece } A_i \text{ en las } n \text{ repeticiones, } i = 1, 2$$

y para enteros no negativos, x_1 y x_2 tales que $x_1 + x_2 = n$ se tiene que:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2) = \binom{n}{x_1, x_2} p_1^{x_1} p_2^{x_2} = \binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n-x_1}$$

es una probabilidad binomial. Obviamente $X_1 + X_2 = n$ así que el vector multinomial es

$$(X_1, X_2) = (X_1, n - X_1)$$

por lo que claramente la segunda componente es redundante y para la primera componente se tiene $X_1 \sim \text{Bi}(n, p_1)$. Por este motivo cuando $n = 2$ el vector multinomial puede identificarse con su primera componente la cuál sigue una distribución binomial. ■

2.5 Propiedades del vector multinomial

2.5.1 Función generatriz de momentos *

La función generatriz de momentos un vector aleatorio $\underline{X} = (X_1, \dots, X_k) \sim \mathcal{M}(n; p_1, \dots, p_k)$ es:

$$M(t_1, t_2, \dots, t_k) = \left(\sum_{i=1}^k p_i e^{t_i} \right)^n.$$

Demostración. Para $\underline{t} \in \mathbb{R}^k$:

$$\begin{aligned} M(\underline{t}) &= \mathbb{E}(\exp\langle \underline{t}, \underline{X} \rangle) = \sum_{\underline{x} \in \text{Sop}(\underline{X})} \exp\langle \underline{t}, \underline{x} \rangle \mathbb{P}(\underline{X} = \underline{x}) \\ &= \sum_{\underline{x} \in \text{Sop}(\underline{X})} \exp\left(\sum_{i=1}^k t_i x_i\right) \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i} \\ &= \sum_{\underline{x} \in \text{Sop}(\underline{X})} \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k (p_i e^{t_i})^{x_i} = \left(\sum_{i=1}^k p_i e^{t_i} \right)^n. \quad \blacksquare \end{aligned}$$

2.5.2 Distribución marginal de las componentes *

La distribución marginal de cada componente X_i es binomial, concretamente, $X_i \sim \text{Bi}(n, p_i)$.

Demostración. Lo haremos para $i = 1$, para el resto de las componentes la prueba es similar. La función generatriz de X_1 es:

$$M_{X_1}(t_1) = M_{\underline{X}}(t_1, 0, \dots, 0) = (p_1 e^{t_1} + p_2 + \dots + p_k)^n = (p_1 e^{t_1} + 1 - p_1)^n,$$

por lo tanto $X_1 \sim \text{Bi}(n, p_1)$. ■

2.5.3 Agrupación de componentes *

Sea $\underline{X} = (X_1, X_2, \dots, X_k)' \sim \mathcal{M}(n; p_1, p_2, \dots, p_k)$. Para $r < k$, se verifica:

$$\left(X_1, X_2, \dots, X_r, n - \sum_{j=1}^r X_j \right)' \sim \mathcal{M} \left(n; p_1, \dots, p_r, 1 - \sum_{j=1}^r p_j \right).$$

Demostración. Obsérvese que $n - \sum_{j=1}^r X_j = \sum_{j=r+1}^n X_j$. La función generatriz de momentos de $(X_1, X_2, \dots, X_r, n - \sum_{j=1}^r X_j)$ es, para $(t_1, \dots, t_{r+1})' \in \mathbb{R}^{r+1}$:

$$\begin{aligned} & \mathbb{E} \left(\exp \left(t_1 X_1 + \dots + t_r X_r + t_{r+1} \left(n - \sum_{j=1}^r X_j \right) \right) \right) \\ &= \mathbb{E} \left(\exp \left(t_1 X_1 + \dots + t_r X_r + t_{r+1} \left(\sum_{j=r+1}^k X_j \right) \right) \right) \\ &= M_{\underline{X}}(t_1, \dots, t_r, \overbrace{t_{r+1}, \dots, t_{r+1}}^{k-r \text{ veces}}) = (p_1 e^{t_1} + \dots + p_r e^{t_r} + (p_{r+1} + \dots + p_k) e^{t_{r+1}})^n \\ &= \left(p_1 e^{t_1} + \dots + p_r e^{t_r} + \left(1 - \sum_{j=r+1}^k p_j \right) e^{t_{r+1}} \right)^n, \end{aligned}$$

por lo tanto $(X_1, X_2, \dots, X_r, n - \sum_{j=1}^r X_j) \sim \mathcal{M} \left(n; p_1, \dots, p_r, 1 - \sum_{j=1}^r p_j \right)$. ■

«¿Y aquí qué estamos haciendo exactamente, lo mismo pero una especie de pseudo-marginal hasta r ? O sea, lo que entiendo es que perdemos granularidad, simplemente clasificamos en esas r cajas y el resto de $k - r$ cajas las metemos en una sola. Y resulta que esa pseudo-marginal (que no sé si posee algún nombre riguroso) sigue siendo una multinomial.»

El nombre riguroso en teoría de la probabilidad es la distribución marginal conjunta del subvector $(X_1, \dots, X_r)$. En teoría de la medida y procesos estocásticos, a esta operación geométrica se la conoce formalmente como un agrupamiento de estados (*state lumping*) o empuje de la medida (*pushforward measure*) mediante una función de partición.

En física, concretamente en termodinámica y mecánica estadística, lo que acabas de describir como “perder granularidad” es posiblemente la operación más importante para entender el universo: se llama *Coarse-Graining* (técnica de grano grueso) o Renormalización.

Imagina un gas. Las k cajas originales son los microestados cuánticos hiperprecisos en los que puede estar una partícula. Sin embargo, tus instrumentos de medición (o tu interés analítico) solo tienen resolución suficiente para distinguir los primeros r niveles de energía. Cualquier partícula que caiga en los niveles desde $r + 1$ hasta k hace que tu máquina encienda una única luz genérica que dice: “Energía Alta”.

Al agrupar esas $k - r$ cajas en una sola, estás pasando de observar microestados a observar un macroestado. Estás perdiendo entropía (información) a propósito para simplificar el sistema. Dado que las partículas siguen siendo independientes y siguen cayendo en “cajas” (solo que ahora una de las cajas es físicamente más ancha porque engloba a varias), la naturaleza fundamental del sistema no cambia. Las leyes de la física (la distribución multinomial) se conservan a escala macroscópica.

Observemos la FGM, pero evaluémosla en $(t_1, \dots, t_r, t_{r+1}, \dots, t_{r+1})$. Aquí es donde reside la brillantez formal (el “truco”). En la demostración de la Binomial, el autor apagaba las

dimensiones haciendo que las t valieran 0. Aquí, el autor no las apaga. Lo que hace es obligar a que las variables $t_{r+1}, t_{r+2}, \dots, t_k$ sean todas exactamente iguales a una nueva variable t_{r+1} .

¿Qué significa esto en el espacio dual (el espacio de las direcciones de exploración)? Estás restringiendo tu hiper-superficie generatriz a un subespacio diagonal. Físicamente, significa que cualquier “tensión” o “temperatura” matemática que apliques a tu nueva “caja resto” (t_{r+1}), debe aplicarse de manera idéntica y simultánea a todas las cajas subyacentes que la componen (de la $r + 1$ a la k). Las has soldado matemáticamente.

Observa cómo el álgebra obedece a esta soldadura geométrica. Cuando fuerzas a que todas compartan el mismo exponente $e^{t_{r+1}}$, la expresión se agrupa:

$$p_{r+1}e^{t_{r+1}} + p_{r+2}e^{t_{r+1}} + \dots + p_k e^{t_{r+1}}$$

El factor de deformación $e^{t_{r+1}}$ actúa como un factor común. Lo sacas fuera del paréntesis (una factorización elemental), y lo que queda dentro es la suma pura de las masas probabilísticas:

$$(p_{r+1} + \dots + p_k) e^{t_{r+1}}$$

Y dado que $\sum_{j=1}^k p_j = 1$, la suma de todo lo que hay desde la caja $r + 1$ hasta la k es, por fuerza de la conservación de la medida, $1 - \sum_{j=1}^r p_j$.

2.5.4 Momentos *

Momentos:

$$\begin{aligned} \mathbb{E}X_i &= np_i; \quad \text{Var}(X_i) = np_i(1 - p_i), \\ \text{Cov}(X_i, X_j) &= -np_i p_j, \\ \rho_{i,j} &= - \left(\frac{p_i p_j}{(1 - p_i)(1 - p_j)} \right)^{1/2}. \end{aligned}$$

Demostración.

- Como $X_i \sim \text{Bi}(n, p_i)$ se tiene $\mathbb{E}X_i = np_i$ y $\text{Var}(X_i) = np_i(1 - p_i)$.
- Se tiene:

$$D_{i,j}^{1,1} M_{\underline{X}}(\underline{t}) = n(n-1) \left(\sum_{r=1}^k p_r e^{t_r} \right)^{n-2} p_i p_j e^{t_i} e^{t_j},$$

entonces:

$$\mathbb{E}(X_i X_j) = D_{i,j}^{1,1} M_{\underline{X}}(\underline{0}) = n(n-1) p_i p_j,$$

finalmente:

$$\text{Cov}(X_i, X_j) = \mathbb{E}X_i X_j - \mathbb{E}X_i \mathbb{E}X_j = n(n-1) p_i p_j - n^2 p_i p_j = -np_i p_j.$$

El coeficiente de correlación lineal es:

$$\rho_{i,j} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i) \text{Var}(X_j)}} = \frac{-np_i p_j}{\sqrt{np_i p_j (1 - p_i)(1 - p_j)}} = - \left(\frac{p_i p_j}{(1 - p_i)(1 - p_j)} \right)^{1/2}. \quad \blacksquare$$

2.5.5 Reproductividad *

Si $\underline{X} \sim \mathcal{M}(n; p_1, p_2, \dots, p_k)$ e $\underline{Y} \sim \mathcal{M}(m; p_1, p_2, \dots, p_k)$ son independientes entonces:

$$\underline{Z} = \underline{X} + \underline{Y} \sim \mathcal{M}(n + m; p_1, p_2, \dots, p_k).$$

Demostración. Como $\underline{X}$ y $\underline{Y}$ son independientes, se tiene:

$$M_{\underline{X}+\underline{Y}}(\underline{t}) = M_{\underline{X}}(\underline{t})M_{\underline{Y}}(\underline{t}) = \left(\sum_{i=1}^k p_i e^{t_i} \right)^n \left(\sum_{i=1}^k p_i e^{t_i} \right)^m = \left(\sum_{i=1}^k p_i e^{t_i} \right)^{n+m},$$

entonces:

$$\underline{X} + \underline{Y} \sim \mathcal{M}(n + m; p_1, p_2, \dots, p_k). \quad \blacksquare$$

2.5.6 Distribución condicionada *

Distribuciones condicionadas:

$$X_1 \mid X_2 = x_2 \sim \text{Bi} \left(n - x_2, \frac{p_1}{1 - p_2} \right).$$

Demostración. La función de probabilidad conjunta del vector (X_1, X_2) es:

$$\begin{aligned} \mathbb{P}(X_1 = x_1, X_2 = x_2) &= \mathbb{P} \left(X_1 = x_1, X_2 = x_2, \sum_{i=3}^k X_i = n - x_1 - x_2 \right) \\ &= \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}, \end{aligned}$$

para $x_1, x_2 \geq 0$ y $x_1 + x_2 \leq n$. Entonces:

$$\begin{aligned} \mathbb{P}(X_1 = x_1 \mid X_2 = x_2) &= \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2)}{\mathbb{P}(X_2 = x_2)} \\ &= \frac{\frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}}{\frac{n!}{x_2! (n - x_2)!} p_2^{x_2} (1 - p_2)^{n - x_2}} \\ &= \frac{(n - x_2)!}{x_1! (n - x_1 - x_2)!} \frac{p_1^{x_1} (1 - p_1 - p_2)^{n - x_1 - x_2}}{(1 - p_2)^{n - x_2}} \\ &= \binom{n - x_2}{x_1} \left(\frac{p_1}{1 - p_2} \right)^{x_1} \left(1 - \frac{p_1}{1 - p_2} \right)^{n - x_1 - x_2}. \quad \blacksquare \end{aligned}$$

1. ¿Quiénes son los actores? (La Ontología del Sistema)

Primero, aclaremos qué son X_1 y X_2 . Ni X_1 ni X_2 son distribuciones multinomiales por sí solas.

Existe un único macro-objeto geométrico: el vector aleatorio $\underline{X} = (X_1, X_2, \dots, X_k)$, el cual sí es multinomial.

Piensa en el experimento clásico: tienes n partículas que vas a lanzar, y tienes k cajas (o niveles de energía) donde pueden caer.

- X_1 es simplemente el número de partículas que terminaron en la caja 1.
- X_2 es el número de partículas que terminaron en la caja 2.

Si miras a X_1 ignorando el resto del universo (lo que discutimos en la pregunta anterior sobre proyectar sombras), su distribución marginal es una Binomial, es decir, $X_1 \sim \text{Bi}(n, p_1)$. Lo mismo para $X_2 \sim \text{Bi}(n, p_2)$.

2. La Física del Condicionamiento: El Colapso de la Realidad

Alguien ha abierto la caja 2, ha contado las partículas dentro, y te ha dicho: “Es un hecho irrefutable que exactamente x_2 partículas están aquí encerradas”.

A partir de este instante, el universo ha cambiado. Se aplican dos leyes físicas brutales:

A. La Conservación de la Masa (El nuevo número de ensayos):

Originalmente tenías n partículas. Si x_2 están definitivamente atrapadas en la caja 2, ¿cuántas partículas quedan libres interactuando en el resto del universo? Obviamente, quedan $n - x_2$ partículas.

Por lo tanto, la variable X_1 ya no compete por n partículas. Su nuevo límite máximo absoluto es $n - x_2$. El “tamaño” de tu sistema físico se ha encogido.

B. La Renormalización Topológica (La nueva probabilidad):

Originalmente, cualquier partícula tenía una probabilidad p_1 de caer en la caja 1, y una probabilidad p_2 de caer en la caja 2.

Pero ahora, para las $n - x_2$ partículas restantes, la caja 2 ha dejado de existir. Es un agujero negro que ya ha tragado su cuota y se ha cerrado.

En teoría de la medida, el espacio de probabilidad completo debe sumar 1. Si eliminas el trozo de espacio correspondiente a p_2 , tu espacio restante tiene una “masa” de $1 - p_2$. Las probabilidades de caer en las cajas restantes ($p_1, p_3, \dots, p_k$) ya no suman 1, suman $1 - p_2$.

Para restaurar la geometría euclídea de tu espacio de probabilidad (para que vuelva a sumar 1), tienes que renormalizar. Debes dilatar (estirar) las probabilidades restantes dividiéndolas por el volumen que queda. Así, la nueva probabilidad efectiva de que una de las partículas restantes caiga en la caja 1 es $\frac{p_1}{1-p_2}$.

El autor quiere saber la probabilidad conjunta de las cajas 1 y 2. ¿Qué hace con las cajas 3, 4, hasta k ? Las colapsa todas en una sola “caja resto”.

Fíjate en el término: $\sum_{j=3}^k X_j = n - x_1 - x_2$.

Esto convierte tu problema multinomial de k dimensiones en un problema de solo 3 dimensiones.

2.5.7 Curva de regresión *

Curva de regresión:

$$\mathbb{E}(X_1 | X_2) = (n - X_2) \frac{p_1}{1 - p_2}$$

que es lineal y por lo tanto coincide con la recta de regresión. Además:

$$\theta_{X_1|X_2}^2 = \rho^2 = \frac{p_1 p_2}{(1 - p_1)(1 - p_2)}.$$

Demostración. Como $X_1 | X_2 = x_2 \sim \text{Bi}\left(n - x_2, \frac{p_1}{1 - p_2}\right)$ se tiene

$$\mathbb{E}(X_1 | X_2 = x_2) = (n - x_2) \frac{p_1}{1 - p_2}. \quad \blacksquare$$

Distribución multivariante

3.1 Vector normal multivariante

Sea un vector aleatorio $\underline{Z} = (Z_1, \dots, Z_n)'$, con $Z_i, i = 1, \dots, n$, variables aleatorias independientes e idénticamente distribuidas (iid) todas ellas con distribución $N(0, 1)$. La fdd conjunta de este vector es, debido a la independencia:

$$f_{\underline{Z}}(\underline{z}) = \prod_{i=1}^n f_{Z_i}(z_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\underline{z}'\underline{z}\right),$$

donde $\underline{z} = (z_1, \dots, z_n)' \in \mathbb{R}^n$. El vector aleatorio $\underline{Z}$ se denomina vector aleatorio normal n -dimensional standard.

3.1.1 Definición

Sea $\underline{Z}$ el vector aleatorio definido anteriormente. Sean $\underline{\mu} \in \mathbb{R}^n$ y A una matriz $n \times n$ de constantes. Entonces el vector aleatorio $\underline{X} = \underline{\mu} + A\underline{Z}$ es un vector aleatorio normal n -dimensional.

3.2 Propiedades del vector normal multivariante

3.2.1 Función generatriz de momentos *

La función generatriz de momentos de un vector aleatorio normal n -dimensional es:

$$M_{\underline{X}}(\underline{t}) = \exp\left(\underline{t}'\underline{\mu} + \frac{1}{2}\underline{t}'\Sigma\underline{t}\right), \quad \underline{t} \in \mathbb{R}^n$$

con $\Sigma = AA'$.

Demostración. En primer lugar determinaremos la función generatriz de momentos del vector $\underline{Z}$:

$$\begin{aligned} M_{\underline{Z}}(\underline{t}) &= \mathbb{E}(\exp\langle \underline{t}, \underline{Z} \rangle) = \mathbb{E} \left(\exp \left(\sum_{i=1}^n t_i Z_i \right) \right) = \prod_{i=1}^n \mathbb{E}(\exp(t_i Z_i)) \\ &= \prod_{i=1}^n M_{Z_i}(t_i) = \prod_{i=1}^n \exp \left(\frac{1}{2} t_i^2 \right) = \exp \left(\frac{1}{2} \sum_{i=1}^n t_i^2 \right) = \exp \left(\frac{1}{2} \underline{t}' \underline{t} \right), \end{aligned}$$

entonces:

$$\begin{aligned} M_{\underline{X}}(\underline{t}) &= \mathbb{E}(\exp\langle \underline{t}, \underline{X} \rangle) = \mathbb{E}(\exp\langle \underline{t}, \underline{\mu} + A\underline{Z} \rangle) = \mathbb{E}(\exp(\underline{t}'(\underline{\mu} + A\underline{Z}))) \\ &= \exp(\underline{t}'\underline{\mu}) \mathbb{E}(\exp((A'\underline{t})'\underline{Z})) = \exp(\underline{t}'\underline{\mu}) M_{\underline{Z}}(A'\underline{t}) \\ &= \exp(\underline{t}'\underline{\mu}) \exp \left(\frac{1}{2} (A'\underline{t})' (A'\underline{t}) \right) \\ &= \exp \left(\underline{t}'\underline{\mu} + \frac{1}{2} \underline{t}' A A' \underline{t} \right) = \exp \left(\underline{t}'\underline{\mu} + \frac{1}{2} \underline{t}' \Sigma \underline{t} \right), \quad \underline{t} \in \mathbb{R}^n. \quad \blacksquare \end{aligned}$$

3.2.2 Media y matriz de covarianzas *

$\mathbb{E}\underline{X} = \underline{\mu}$ y $\text{Var } \underline{X} = \Sigma$. Estos dos parámetros caracterizan perfectamente a una distribución normal n -dimensional. Por lo tanto en lo siguiente se denotará $\underline{X} \sim N_n(\underline{\mu}, \Sigma)$.

Demostración. Se tiene que $\mathbb{E}\underline{Z} = \underline{0}$ y $\text{Var}(\underline{Z}) = I_n$, entonces:

$$\begin{aligned} \mathbb{E}\underline{X} &= \mathbb{E}(\underline{\mu} + A\underline{Z}) = \underline{\mu} + A\mathbb{E}\underline{Z} = \underline{\mu}; \\ \text{Var } \underline{X} &= \text{Var}(\underline{\mu} + A\underline{Z}) = \text{Var}(A\underline{Z}) = A \text{Var}(\underline{Z}) A' = A A' = \Sigma. \end{aligned}$$

$\underline{\mu}$ y Σ caracterizan a la distribución pues son los únicos parámetros necesarios para determinar completamente la función generatriz de momentos. ■

3.2.3 Caso no singular: función de densidad *

Si $\det \Sigma \neq 0$, entonces el vector aleatorio $\underline{X} \sim N_n(\underline{\mu}, \Sigma)$ es absolutamente continuo con fdd:

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2} (\det \Sigma)^{1/2}} \exp \left(-\frac{1}{2} (\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) \right), \quad \underline{x} \in \mathbb{R}^n.$$

Demostración. La demostración consiste en aplicar el Teorema del cambio de variables. Sea $\underline{Z} = (Z_1, \dots, Z_n)$ un vector aleatorio con fdd conjunta:

$$f_{\underline{Z}}(\underline{z}) = \frac{1}{(2\pi)^{n/2}} \exp \left(-\frac{1}{2} \underline{z}' \underline{z} \right)$$

y consideremos la función (cambio de variables):

$$g : \underline{z} \in \mathbb{R}^n \mapsto \underline{x} = \underline{\mu} + A\underline{z}.$$

Obsérvese que $\det \Sigma = \det(A'A) = \det(A)^2 \neq 0$, por lo que $\det(A) \neq 0$, luego A es invertible y por tanto existe la función inversa de g :

$$g^{-1} : \underline{x} \in \mathbb{R}^n \mapsto \underline{z} = A^{-1}(\underline{x} - \underline{\mu})$$

la cuál es diferenciable con:

$$\text{abs}(\det J_{g^{-1}}) = \text{abs}(\det A^{-1}) = \text{abs}(1/\det(A)) = \text{abs}(1/(\det \Sigma)^{1/2})$$

y finalmente aplicando el Teorema del cambio de variables obtenemos:

$$\begin{aligned} f_{\underline{X}}(\underline{x}) &= f_{\underline{Z}}(g^{-1}(\underline{x})) \text{abs}(\det J_{g^{-1}}) \\ &= \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp \left(-\frac{1}{2}(\underline{x} - \underline{\mu})'(A^{-1})'A^{-1}(\underline{x} - \underline{\mu}) \right) \\ &= \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp \left(-\frac{1}{2}(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu}) \right). \quad \blacksquare \end{aligned}$$

3.2.4 Caso bidimensional *

En particular para el caso bidimensional la fdd conjunta es para $(x_1, x_2) \in \mathbb{R}^2$:

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{(2\pi)\sigma_1\sigma_2(1-\rho^2)^{1/2}} \\ &\times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right) \right\} \\ &= \frac{1}{(2\pi)\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} (z_1^2 + z_2^2 - 2\rho z_1 z_2) \right\}, \end{aligned}$$

donde ρ es el coeficiente de correlación lineal entre X_1 y X_2 y $z_i = (x_i - \mu_i)/\sigma_i$, $i = 1, 2$. En tal caso, se denota usualmente:

$$(X_1, X_2)' \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho).$$

Demostración. La demostración consiste en substituir en la expresión general de la fdd los valores:

$$\underline{\mu} = (\mu_1, \mu_2)' \quad \text{y} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

donde $\mu_i = \mathbb{E}X_i$, $\sigma_i^2 = \text{Var}(X_i)$, $i = 1, 2$ y $\sigma_{12} = \text{Cov}(X_1, X_2)$. Si ρ es el coeficiente de correlación lineal, se tiene $\sigma_{12} = \rho\sigma_1\sigma_2$. Se tiene:

$$\begin{aligned} \det \Sigma &= \sigma_1^2 \sigma_2^2 (1 - \rho^2) \\ \Sigma^{-1} &= \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}. \quad \blacksquare \end{aligned}$$

Conexión con la Geometría Diferencial: El Tensor Métrico

Aquí es donde la probabilidad se da la mano con la Teoría de la Relatividad General de Einstein. Observa el término dentro de la exponencial del caso general: $(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu})$.

En geometría euclidiana clásica (teorema de Pitágoras), la distancia al cuadrado entre dos puntos se mide como $\Delta x^2 + \Delta y^2$. Pero si el espacio está curvado o deformado, necesitamos un tensor métrico, denotado usualmente como g_{ij} , para medir distancias reales: $\sum g_{ij}\Delta x_i \Delta x_j$.

En la teoría de la probabilidad, la inversa de la matriz de covarianzas, Σ^{-1} , actúa exactamente como el tensor métrico del espacio de probabilidades. Define una geometría no euclidiana donde la "distancia" no se mide en metros con una regla recta, sino en unidades

de desviación estándar, respetando la correlación entre variables. Esta métrica brutalmente elegante se conoce como la Distancia de Mahalanobis.

La expansión algebraica artesanal que ves en la imagen para el caso $n = 2$ no es más que el acto de escribir las ecuaciones de la métrica del espacio-tiempo probabilístico, bajando del Olimpo de las formas bilineales sin coordenadas al barro humano de los sistemas de referencia cartesianos (las variables x_1 y x_2).

3.2.5 Caso singular *

Si $\det \Sigma = 0$, entonces el vector aleatorio $\underline{X} \sim N_n(\underline{\mu}, \Sigma)$ es singular y por lo tanto no tiene fdd.

Demostración. El vector aleatorio $\underline{X}$ es la transformación del vector aleatorio $\underline{Z}$ mediante la transformación afín:

$$g(\underline{z}) = \underline{\mu} + A\underline{z}.$$

Si $\det(\Sigma) = \det(A'A) = 0$, entonces $\det(A) = 0$ y $r = \text{rango}(A) < n$, entonces la imagen de la transformación es una variedad afín, V , cuya dimensión coincide con $r < n$ y por tanto $\mu_n(V) = 0$. El conjunto V verifica:

- $\mathbb{P}(\underline{X} \in V) = \mathbb{P}(\underline{Z} \in \mathbb{R}^n) = 1$,
- $\mu_n(V) = 0$,
- Puede probarse que la distribución de $\underline{X}$ no tiene átomos, es decir el vector $\underline{X}$ es singular. ■

3.2.6 Distribución de los subvectores *

Si $\underline{X} = (X_1, \dots, X_n)' \sim N_n(\underline{\mu}, \Sigma)$, entonces $X_i \sim N_1(\mu_i, \sigma_i^2)$ con $\sigma_i^2 = \text{diag}_i(\Sigma)$. De forma más general cualquier subvector de un vector normal n -dimensional es normal multidimensional. Se tiene pues si $\underline{X} = (\underline{X}_1', \underline{X}_2')'$ con $\underline{X}_1$ un vector aleatorio p -dimensional:

$$\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim N_n \left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_2 \end{pmatrix} \right)$$

entonces:

$$\underline{X}_1 \sim N_p(\underline{\mu}_1, \Sigma_1).$$

Demostración. La fgm de $\underline{X} = (\underline{X}_1', \underline{X}_2')'$ para $\underline{t} = (t_1', t_2')'$ es:

$$M_{\underline{X}}(\underline{t}) = \exp \left(\underline{t}' \underline{\mu} + \frac{1}{2} \underline{t}' \Sigma \underline{t} \right),$$

entonces la fgm del subvector $\underline{X}_1$ es:

$$\begin{aligned} M_{\underline{X}} \begin{pmatrix} t_1 \\ 0 \end{pmatrix} &= \exp \left\{ (t_1', 0') \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix} + \frac{1}{2} (t_1', 0') \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_2 \end{pmatrix} \begin{pmatrix} t_1 \\ 0 \end{pmatrix} \right\} \\ &= \exp \left(t_1' \underline{\mu}_1 + \frac{1}{2} t_1' \Sigma_1 t_1 \right), \end{aligned}$$

de donde:

$$\underline{X}_1 \sim N_p(\underline{\mu}_1, \Sigma_1). \quad \blacksquare$$

3.2.7 Incorrelación implica independencia *

Si $\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_n \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \right)$ entonces X_1 y X_2 son vectores normales multidimensionales independientes. En particular si $\underline{X} = (X_1, \dots, X_n)' \sim N_n(\underline{\mu}, \Sigma)$, y Σ es una matriz diagonal (es decir todas las covarianzas son nulas), entonces $X_1, \dots, X_n$ son variables normales independientes. Por lo tanto, bajo normalidad multivariante, la incorrelación implica independencia.

Demostración. La independencia se probará mediante la factorización de la fgm de $\underline{X}$. Sea $\underline{t} = (t'_1, t'_2)'$, entonces:

$$\begin{aligned} M_{\underline{X}} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} &= \exp \left\{ (t'_1, t'_2) \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \frac{1}{2} (t'_1, t'_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \right\} \\ &= \exp \left(t'_1 \mu_1 + t'_2 \mu_2 + \frac{1}{2} t'_1 \Sigma_1 t_1 + \frac{1}{2} t'_2 \Sigma_2 t_2 \right) \\ &= \exp \left(t'_1 \mu_1 + \frac{1}{2} t'_1 \Sigma_1 t_1 \right) \exp \left(t'_2 \mu_2 + \frac{1}{2} t'_2 \Sigma_2 t_2 \right) \\ &= M_{X_1}(t_1) M_{X_2}(t_2). \quad \blacksquare \end{aligned}$$

3.2.8 Distribuciones condicionadas *

Las distribuciones condicionadas de las componentes de una distribución normal multivariante son también normales multivariantes. En particular para el caso bidimensional: Si $(X_1, X_2)' \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ no degenerada, entonces la distribución condicionada de X_1 a $X_2 = x_2$ es:

$$N \left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2), \sigma_1^2 (1 - \rho^2) \right).$$

Demostración. Sea $z_i = (x_i - \mu_i) / \sigma_i$, $i = 1, 2$. Se tiene que $X_2 \sim N(\mu_2, \sigma_2^2)$, entonces

$$\begin{aligned} f_{X_1|X_2=x_2}(x_1) &= \frac{f(x_1, x_2)}{f_{X_2}(x_2)} \\ &= \frac{\frac{1}{(2\pi)\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} (z_1^2 + z_2^2 - 2\rho z_1 z_2) \right\}}{\frac{1}{\sqrt{2\pi}\sigma_2} \exp \left(-\frac{1}{2} z_2^2 \right)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1(1-\rho^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} (z_1^2 + \rho^2 z_2^2 - 2\rho z_1 z_2) \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1(1-\rho^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} (z_1 - \rho z_2)^2 \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1(1-\rho^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho^2)\sigma_1^2} \left(x_1 - \mu_1 - \frac{\rho\sigma_1}{\sigma_2} (x_2 - \mu_2) \right)^2 \right\}, \end{aligned}$$

que es la fdd de una distribución $N \left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2), \sigma_1^2 (1 - \rho^2) \right)$. ■

3.2.9 Curvas de regresión *

También en el caso bidimensional (no singular), las curvas de regresión son lineales:

$$\mathbb{E}(X_1 | X_2) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (X_2 - \mu_2)$$

$$\mathbb{E}(X_2 | X_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (X_1 - \mu_1)$$

Por lo tanto la curvas de regresión coinciden con las rectas de regresión y

$$\theta_{X_1|X_2}^2 = \theta_{X_2|X_1}^2 = \rho^2.$$

Demostración. Como $X_1 | X_2 = x_2 \sim N\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2), \sigma_1^2 (1 - \rho^2)\right)$ se tiene

$$\mathbb{E}(X_1 | X_2 = x_2) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2),$$

es decir la curva de regresión es lineal y entonces $\rho^2 = \theta_{X_1|X_2}^2$. ■

3.3 Intuición

La intuición física: La gota de tinta en un cristal anisotrópico

Dejemos el álgebra a un lado y visualicemos esto en nuestra mente. Imagina que tienes un vector normal estándar $\underline{Z}$. Físicamente, imagina que dejas caer una gota de tinta luminiscente en un estanque de agua en perfecto reposo. La tinta se difunde formando una esfera perfecta (o un círculo perfecto en 2D). La probabilidad de encontrar una partícula de tinta decae a medida que te alejas del centro en cualquier dirección por igual. El espacio es isotrópico (igual en todas direcciones).

Ahora, imagina que dejas caer esa misma tinta dentro de un cristal o un fluido donde las corrientes fluyen más rápido en una diagonal que en la otra. ¿Qué le pasa a la mancha de tinta? Ya no es un círculo perfecto; se estira, se deforma y se rota, convirtiéndose en una elipse.

La matriz A es exactamente el operador matemático que aplica esa deformación y rotación al espacio físico. La correlación ρ mide cuán estirada está esa elipse en una dirección diagonal. Si $\rho = 0$, no hay cizalladura diagonal, y volvemos a tener ejes ortogonales (independencia).

(Distancia de Mahalanobis y el Tensor Inverso):

Para un vector $\underline{X} \sim N_n(\underline{\mu}, \Sigma)$ absolutamente continuo ($\det \Sigma > 0$), el lugar geométrico de equiprobabilidad en el hiperespacio se define por las superficies donde el argumento de su exponencial es constante:

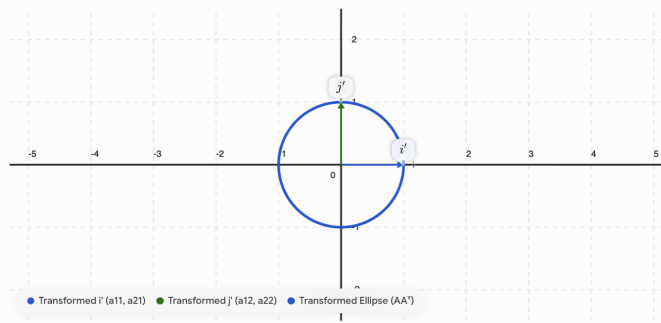
$$d^2(\underline{x}, \underline{\mu}) = (\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) = c^2$$

Teorema 3 (El Dualismo del Esfuerzo Cortante):

El tensor Σ^{-1} induce una métrica Riemanniana euclídea en el espacio muestral. Los autovectores de Σ (las columnas de nuestra matriz de cizalladura Q) dictan ortogonalmente las direcciones invariantes de la transformación de deformación. Los autovalores λ_i de Σ representan las varianzas a lo largo de esas direcciones principales.

Conexión Final: Cuando mides la correlación entre dos variables (un fenómeno estadístico), lo que estás haciendo geométricamente es proyectar el tensor de deformación multidimensional Σ sobre un plano 2D bidimensional, observando el “esfuerzo cortante” residual que dejó la matriz de rotación Q al desalinearse la realidad respecto a los ejes de tu laboratorio.

Transformation Matrix & Covariance



Σ_{11}
1.00

Σ_{12}
0.00

Σ_{22}
1.00

a11

1

a12

0

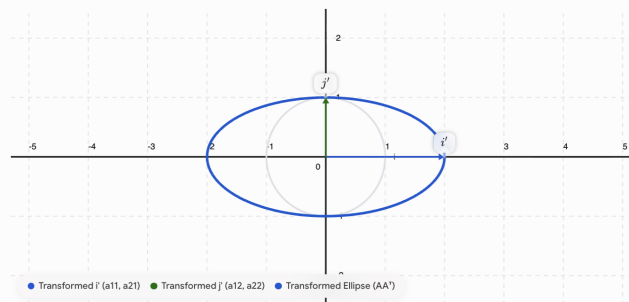
a21

0

a22

1

Transformation Matrix & Covariance



Σ_{11}
4.00

Σ_{12}
0.00

Σ_{22}
1.00

a11

2

a12

0

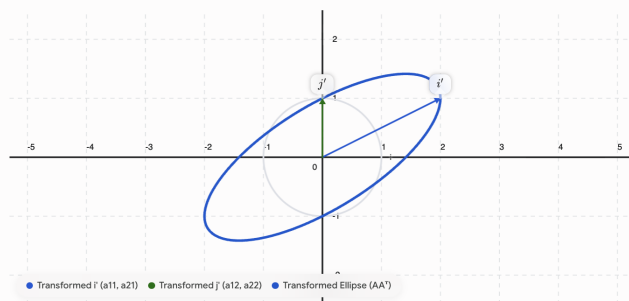
a21

0

a22

1

Transformation Matrix & Covariance



Σ_{11}
4.00

Σ_{12}
2.00

Σ_{22}
2.00

a11

2

a12

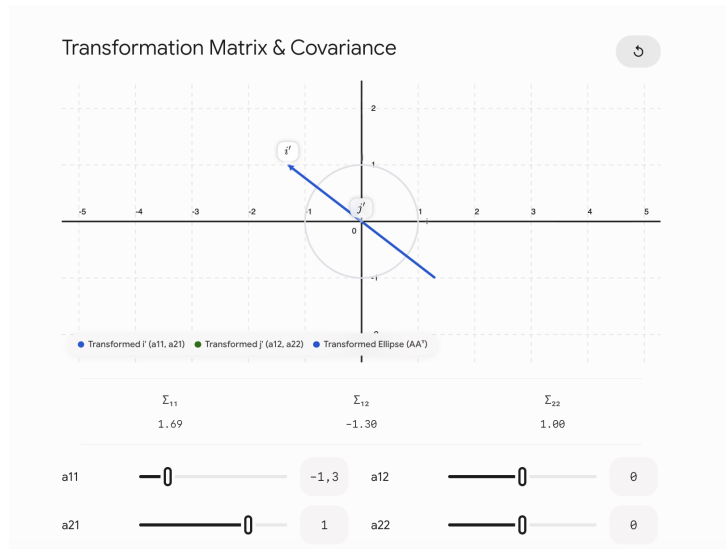
0

a21

1

a22

1



«Supongamos una distribución normal bivalente, con una matriz Σ de varianzas. Mi X_1 vive en el espacio de Hilbert L^2 , ídem con X_2 . Pero eso nos da igual ahora porque lo que nos importa es que X_1 y X_2 lo que hacen es formar una distribución de probabilidad en $\mathbb{R}^2$, que por ejemplo puede verse como absolutamente continua si Σ es regular. Representamos $\mathbb{R}^2$ pero usando como ejes coordenados X_1 y X_2 . Pero esto es una particularidad. Esto nunca lo había pensado, es como que en cierto sentido es una forma concreta de darle coordenadas a $\mathbb{R}^2$ el verlas en términos de X_1 y X_2 , ¿no? El caso es que a cada punto de ese $\mathbb{R}^2$ con coordenadas así definidas posee su densidad. (¿Ese $\mathbb{R}^2$ se llama espacio muestral? ¿O cómo?). Si Σ es diagonal, entonces resulta que las multivariantes forman elipses (círculos en concreto si las varianzas son la misma) con ejes principales los ejes coordenados X_1 y X_2 . Pero si Σ no es diagonal, entonces resulta que las multivariantes forman elipses cuyos ejes principales son diferentes a los ejes. Pero si yo tomara por así decir unas variables aleatorias X_3 y X_4 en esos ejes, ¡la distribución sería independiente! Esto es un verdadero lío para mí. Porque no sé qué quiere decir “diagonalizar” este espacio $\mathbb{R}^2$ en concreto. Veo grandes diferencias con respecto a lo típico que yo sé de diagonalizar una aplicación lineal en $\mathbb{R}^2$, donde encuentro los autovectores de la matriz; esto es mucho más confuso, porque naturalmente esas variables X_3 y X_4 están “en la dirección” de esos autovectores, pero no tiene ningún sentido para mí decir que una variable aleatoria real está asociada a una dirección, no sé. Por otro lado, ya sé que Σ es real y simétrica y por tanto tiene autovalores reales, es diagonalizable; ese motivo entiendo que hace que este análisis se pueda hacer de ese modo, y de nuevo, es igual que por ejemplo el elipsoide al que da lugar el tensor de inercia de un sólido rígido. Mmm... El caso es que aquí simplemente lo que cambia es que la distancia de Mahalanobis induce lugares de “distancia constante” respecto al origen usando por medio ese tensor métrico que es Σ^{-1} ... No sé»

I. La Anatomía del Espacio: ¿Dónde vivimos y dónde proyectamos?

Tu primera gran crisis existencial (y muy acertada) es: ¿Qué es ese $\mathbb{R}^2$? ¿Es el espacio muestral?

Aquí necesitamos rigor topológico. En probabilidad formal, partimos de un espacio abstracto $(\Omega, \mathcal{F}, \mathbb{P})$.

El Espacio Muestral es Ω . Aquí viven los “eventos puros del universo”. No tiene por qué tener métrica, ni coordenadas, ni ser $\mathbb{R}^2$. Puede ser el conjunto de todas las caras de un dado o los estados cuánticos de un electrón.

Tus variables aleatorias X_1 y X_2 (que como bien dices, si tienen varianza finita, habitan en el espacio de Hilbert $L^2(\Omega, \mathbb{P})$) son funciones que viajan desde ese universo abstracto hacia el mundo de los números reales: $X_i : \Omega \rightarrow \mathbb{R}$.

Cuando agrupas X_1 y X_2 en un vector aleatorio $\underline{X} = (X_1, X_2)'$, lo que haces es construir un puente medible desde Ω hacia $\mathbb{R}^2$. La genialidad aquí es que este vector $\underline{X}$ induce una nueva medida de probabilidad.

Intuición Geométrica Pausada: Imagina que Ω es una caverna oscura e inescrutable (al estilo de Platón). Las variables aleatorias X_1 y X_2 son focos de luz que proyectan sombras de lo que ocurre en Ω sobre una pared plana. Esa pared plana es $\mathbb{R}^2$.

Ese $\mathbb{R}^2$ NO es el espacio muestral. Formalmente se llama Espacio de Estados, Espacio Inducido o Rango. Y el espacio de probabilidad inducido por el vector aleatorio es exactamente $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_{\underline{X}})$. Le damos coordenadas cartesianas (x_1, x_2) simplemente porque hemos decidido usar las variables X_1 y X_2 como nuestros ejes de referencia para observar el fenómeno.

II. El Misterio de la “Dirección” de una Variable Aleatoria

Llegamos a tu duda más aguda: «no tiene ningún sentido para mí decir que una variable aleatoria real está asociada a una dirección».

Tienes toda la razón en que una variable aleatoria pura es solo un mapeo a $\mathbb{R}$. No tiene “dirección” per se. Pero recuerda que tu vector $\underline{X}$ ya ha mapeado el universo a ese plano $\mathbb{R}^2$.

Si estás en ese plano $\mathbb{R}^2$ con ejes (x_1, x_2) , cualquier combinación lineal de tus variables aleatorias originales representa geoméricamente la proyección del vector aleatorio sobre una nueva dirección en el plano.

Definición Formal (Proyección Aleatoria): Sea $\vec{v} = (v_1, v_2)' \in \mathbb{R}^2$ un vector unitario (una dirección geométrica pura). Si realizamos el producto escalar estándar entre esta dirección estática y nuestro vector aleatorio $\underline{X}$, creamos una nueva variable aleatoria unidimensional Y :

$$Y = \langle \vec{v}, \underline{X} \rangle = v_1 X_1 + v_2 X_2$$

Aquí se resuelve tu paradoja: una variable aleatoria no “es” una dirección, pero en el contexto multivariante, cada dirección espacial $\vec{v}$ en el espacio de estados $\mathbb{R}^2$ engendra una variable aleatoria específica al proyectar el vector $\underline{X}$ sobre ella. Explorar $\mathbb{R}^2$ rotando tus ejes equivale a explorar todas las posibles variables aleatorias que se pueden formar sumando linealmente X_1 y X_2 .

III. La Diagonalización y la Metáfora del Sólido Rígido

Tocas el corazón del Teorema Espectral al mencionar el tensor de inercia. Es la analogía perfecta.

En física mecánica, si tomas una patata (un sólido amorfo) y tratas de hacerla girar sobre un eje arbitrario (ejes X e Y de tu laboratorio), la patata cabeceará y vibrará violentamente. Esto ocurre porque el tensor de inercia no es diagonal; existen productos de inercia que acoplan el movimiento entre los ejes. El Teorema Espectral te dice: “Gira tu sistema de coordenadas del laboratorio. Existen tres ejes ortogonales mágicos (autovectores) incrustados

en la patata. Si la haces girar por ellos, girará perfectamente suave, porque los productos de inercia se anulan (se diagonaliza la matriz)”.

En probabilidad ocurre exactamente lo mismo, pero cambiamos masa por probabilidad y el tensor de inercia por la matriz de varianzas y covarianzas Σ .

El Teorema de Diagonalización Probabilística: La matriz Σ contiene varianzas en la diagonal y covarianzas fuera de ella. La covarianza mide cómo dos variables “cabecean” juntas, es decir, su grado de acoplamiento lineal. Si rotas tus ejes espaciales en $\mathbb{R}^2$, estás definiendo nuevas variables aleatorias (como hemos visto en la Sección II).

“Diagonalizar el espacio $\mathbb{R}^2$ ” significa encontrar un vector de rotación ortogonal (una matriz de cambio de base Q) que reoriente tus ejes geométricos para que apunten en la dirección de los autovectores de Σ . Llamemos a estas nuevas direcciones $\vec{v}_1$ y $\vec{v}_2$. Las variables aleatorias proyectadas sobre estos ejes son $X_3 = \vec{v}_1^T \underline{X}$ y $X_4 = \vec{v}_2^T \underline{X}$.

Como $\vec{v}_1$ y $\vec{v}_2$ son autovectores ortogonales de una matriz simétrica, el “producto de inercia” probabilístico (la covarianza) entre X_3 y X_4 se vuelve cero. Las variables son incorreladas.

Y aquí viene la magia suprema del espacio normal: bajo normalidad multivariante, la incorrelación implica independencia. Al encontrar los autovectores, has desenredado topológicamente la realidad. Las elipses rotadas ahora están alineadas perfectamente con tu nueva cuadrícula, y el sistema hipercomplejo se puede tratar como dos problemas unidimensionales totalmente separados y puros.

IV. Mahalanobis: Cuando la Varianza Dobla el Espacio

Tu intuición final es deslumbrante: «la distancia de Mahalanobis induce lugares de ‘distancia constante’ respecto al origen usando por medio ese tensor métrico que es Σ^{-1} ».

Has descubierto la Geometría Riemanniana que subyace en la probabilidad. En la función de densidad de la normal multivariante, el exponente es:

$$-\frac{1}{2}(\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu})$$

Si el universo fuera isotrópico (esferas en lugar de elipses, varianzas de 1 y covarianzas de 0), Σ^{-1} sería la matriz Identidad I . La expresión colapsaría a la norma euclídea clásica: $x_1^2 + x_2^2$.

Pero al introducir correlaciones y varianzas distintas, el “espacio” probabilístico se deforma. La matriz inversa Σ^{-1} actúa literalmente como el tensor métrico g_{ij} de la relatividad general o de la topología diferencial.

- Euclides mide con una regla rígida de acero.
- Mahalanobis mide con una regla elástica que se estira y encoge dependiendo de hacia dónde estés apuntando, dictada por la incertidumbre (Σ).

Un punto que euclidianamente está muy “lejos” del centro puede estar probabilísticamente “cerca” si se encuentra a lo largo del semieje mayor de la elipse de covarianza (dirección de alta varianza). Los contornos de densidad constante (las elipses) son, en esta topología deformada, esferas perfectas, porque todos los puntos en ese borde están a la misma “distancia de Mahalanobis” del centro.

«Pero digo, es muy radical que pueda encontrar independencia entre dos variables simplemente rotando, ¿no?»

Sí, este “milagro” geométrico de lograr independencia puramente rotando el espacio es una exclusividad absoluta y casi divina de la distribución normal multivariante.

Fuera de este ecosistema hiper-específico, rotar el espacio y diagonalizar la matriz de covarianza Σ solo te garantiza que las nuevas variables sean incorreladas (su covarianza cruzada se anula), pero no garantiza que sean independientes. Como dicta la teoría de la probabilidad, la incorrelación no implica independencia en general. Existen infinidad de ejemplos de vectores aleatorios con covarianza cero, pero donde el conocimiento de una variable sigue condicionando brutalmente el resultado de la otra. Sin embargo, bajo el paraguas estructural de la normalidad multivariante, la incorrelación implica matemáticamente la independencia absoluta.

Conexión Analítica: Ecuaciones en Derivadas Parciales

Este fenómeno de “rotar para independizar” no es más que el eco probabilístico de una de las técnicas más formidables del Análisis Matemático y las Ecuaciones en Derivadas Parciales (EDPs): el Método de Separación de Variables.

Cuando los físicos resuelven la Ecuación del Calor de Fourier o la Ecuación de Schrödinger para un oscilador armónico en mecánica cuántica (cuyas soluciones estado base son campanas de Gauss), asumen de partida que la solución multidimensional buscará un equilibrio de la forma $\Psi(x, y) = X(x)Y(y)$.

Si tu placa térmica es cuadrada pero está torcida respecto a tus coordenadas espaciales, el flujo de calor en el eje X interactuará (dependerá) del flujo en el eje Y .

Si “rotas tu laboratorio” para alinear los ejes espaciales con los bordes geométricos de la placa, de repente la frontera del dominio se vuelve separable. El flujo temporal en X y en Y se resuelven como dos universos cerrados paralelos. Has diagonalizado el operador laplaciano. La independencia estadística es simplemente una manifestación de la ortogonalidad funcional.

Conexión Física: Mecánica Cuántica y Observables

En el Análisis Funcional, los estados del universo físico viven en un Espacio de Hilbert complejo. Medir dos propiedades de un sistema (dos “variables aleatorias” cuánticas, como la posición y el momento de una partícula) de manera independiente requiere que los operadores hermíticos que representan esas magnitudes compartan una base ortonormal común de autovectores.

Buscar los autovectores de la matriz de covarianza en probabilidad y rotar tu sistema para “independizar” las variables macroscópicas es filosóficamente idéntico a ejecutar una transformación unitaria en mecánica cuántica. Estás buscando el marco de referencia intrínseco del universo, purgando el acoplamiento espurio que tú introdujiste al elegir una base cartesiana arbitraria al comenzar tu experimento.

I. Geometría de la Proyección: Las Sombras Perfectas (Marginalización)

Imagina un hiper-elipsoide de 100 dimensiones flotando en el espacio. Representa un vector normal de 100 variables acopladas. Si proyectamos una luz sobre ese objeto hiperdimensional para ver su sombra en una pared de 2 dimensiones (es decir, si ignoramos 98 variables y nos quedamos solo con las distribuciones marginales de 2 de ellas), ¿qué forma tiene la sombra?

Si proyectas un cubo 3D sobre una pared, dependiendo del ángulo, la sombra puede ser un cuadrado, un rectángulo o un hexágono. La topología de la sombra de un cubo cambia radicalmente. Sin embargo, la distribución normal posee una simetría tan absoluta que su

sombra siempre conserva su identidad.

Teorema de Marginalización Euclidiana:

Cualquier subvector de un vector normal n -dimensional es, ineludiblemente, un vector normal multidimensional de menor dimensión. Si tienes $\underline{X} = (\underline{X}_1, \underline{X}_2)'$, al extirpar $\underline{X}_2$ integrando su existencia hasta el infinito, el bloque restante $\underline{X}_1$ sigue siendo estrictamente normal con su media $\underline{\mu}_1$ y su sub-matriz de covarianza Σ_1 intactas.

Filosofía Geométrica: La marginalización es el acto de la “ignorancia matemática”. Estás integrando y promediando todas las posibilidades de las dimensiones que no puedes ver. Que la sombra siga siendo perfecta significa que la naturaleza del ruido gaussiano es fractal: el ruido proyectado sigue siendo ruido gaussiano.

«Vamos a ver, yo tengo un sombrero en 3D que es básicamente la fdd sobre el plano en $\mathbb{R}^2$, pues la altura. Y ahora, si yo digo $X_1|X_2 = x_2$, lo que estoy es cortando el sombrero por donde $X_2 = x_2$ (esa intersección) y renormalizando, dividiendo por la densidad de $f_{X_2}(x_2)$. Generalmente, si yo hago eso, lo que hago es proyectar sobre el espacio que depende informacionalmente de X , pero en concreto resulta que tanto para la binomial como la normal multivariante, en concreto, la proyección es sobre el propio plano que forman X y el eje determinista en ese espacio de Hilbert abstracto. Me gustaría ver intuitiva y geoméricamente por qué esto es así, porque podría ser otra cosa. Pero eso entiendo que es verlo en el espacio de Hilbert; pero en el espacio inducido $\mathbb{R}^2$, en cambio, lo que hago es lo que dije, cortar el sombrero por la recta $X_2 = x_2$. Y entonces, según entiendo, el trasladado de esa noción que hemos visto en el espacio de Hilbert aquí es que lo nuevo que obtengo sigue siendo una normal; y como normal que es, en principio, es simplemente como la propia X_2 pero modificada por un cierto factor de dilatación y sumando una cierta constante. Y eso se puede condensar en una función lineal, y por eso sale lo que sale. Pero me falta intuición, de nuevo, de por qué tiene la forma que tiene, exacta, de la μ y de la ρ y de la proporción de σ s, etc., y no otra»

3.3.1 La caja contenedora: la forma exacta de la media y la varianza condicionadas

I. El Mapa Topográfico y la Caja Contenedora

Al mirar el sombrero gaussiano desde arriba, lo que ves son sus curvas de nivel: elipses concéntricas. Cada elipse representa un lugar donde la probabilidad es constante. Vamos a fijarnos en la elipse principal, la “elipse de 1-sigma”, que condensa la geometría del sistema.

Dibuja mentalmente esta elipse en el plano $\mathbb{R}^2$. Ahora, encierra esa elipse en la caja rectangular más pequeña posible (una caja delimitadora o bounding box con lados paralelos a los ejes X_1 y X_2).

- La anchura total de esta caja rígida es exactamente $2\sigma_1$.
- La altura total de esta caja rígida es exactamente $2\sigma_2$.

Si la correlación ρ es 0, la elipse está perfectamente derecha. Toca los bordes de tu caja exactamente en el centro de los lados (en una cruz perfecta vertical y horizontal).

Pero si la correlación $\rho > 0$, la elipse está inclinada en diagonal, de esquina a esquina de tu caja. Ya no toca los bordes de la caja en el medio, sino que se apoya cerca de las esquinas.

II. La Intuición de la Media Condicionada: El Desplazamiento del Centro

Cuando tú dices " $X_1|X_2 = x_2$ ", la acción geométrica que estás haciendo en este mapa 2D es trazar una línea horizontal recta que atraviesa tu caja y tu elipse a la altura x_2 . Esa línea corta a la elipse formando una cuerda horizontal.

La distribución condicionada es simplemente esa cuerda. Su centro será la nueva media condicionada, $\mathbb{E}(X_1|X_2)$, y su longitud dictará la nueva varianza condicionada. ¿Dónde está el centro exacto de esa cuerda horizontal?

Paso 1: La Regla de los Puntos Medios. Si trazas muchas cuerdas horizontales a distintas alturas x_2 y marcas el punto medio de cada una, la geometría clásica de las secciones cónicas nos dice que todos esos puntos medios forman una línea recta perfecta. Esta línea recta geométrica es, literalmente, la curva de regresión, que al ser lineal coincide exactamente con la recta de regresión.

Paso 2: La Proporción del Viaje. Estás situado en el centro del universo (μ_1, μ_2) . Quieres subir a la altura x_2 . La distancia vertical que recorres es $(x_2 - \mu_2)$.

Paso 3: El Factor de Escala Horizontal. Como tu caja tiene una altura regida por σ_2 y una anchura regida por σ_1 , no puedes simplemente sumar "metros de altura" a "litros de anchura". Tienes que convertir tu viaje vertical en un porcentaje: has subido $\frac{x_2 - \mu_2}{\sigma_2}$ unidades estandarizadas.

Paso 4: El Efecto de la Inclinación (ρ). Al subir ese porcentaje, ¿cuánto te arrastra la elipse hacia la derecha? Si $\rho = 1$, la elipse es una línea diagonal perfecta de esquina a esquina de la caja; si subes un 10% de la caja, te arrastra hacia la derecha exactamente un 10% de la caja. Si $\rho = 0$, la elipse no está inclinada; subes un 10% y te arrastra 0% a la derecha. Por tanto, el porcentaje de arrastre horizontal es exactamente dictado por la correlación: $\rho \times (\text{Porcentaje de subida})$.

Paso 5: La Traducción Final. Has sido arrastrado hacia la derecha un factor de $\rho \frac{x_2 - \mu_2}{\sigma_2}$. Para saber cuánto es esto en las unidades físicas reales del eje X_1 , multiplicas por el tamaño de tu semianchura σ_1 .

Conclusión Geométrica de la Media: El centro de tu cuerda (la nueva esperanza) se ha desplazado desde μ_1 exactamente la cantidad $\rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2)$. Es pura proporción de triángulos dentro de una caja dictada por la inclinación topográfica.

III. La Intuición de la Varianza Condicionada: El Aplastamiento

Ahora vamos a ver por qué la fórmula de la varianza es exactamente $\sigma_1^2(1 - \rho^2)$, independientemente de a qué altura x_2 hagas el corte. Fíjate en la longitud de esa cuerda horizontal que acabas de trazar.

La varianza incondicional de X_1 es σ_1^2 . Esto significa que la incertidumbre total que tienes en el eje horizontal es la anchura total de la caja contenedora.

Cuando la elipse se inclina ($\rho > 0$), se tiene que estirar hacia las esquinas para seguir tocando las paredes de la caja (porque la varianza total de la sombra no puede cambiar). Al estirarse hacia las esquinas sin salirse de la caja, la elipse se vuelve forzosamente más estrecha o afilada a lo largo de cualquier corte horizontal.

La Constricción: Al condicionar, tú ya no deambulas por toda la caja; estás atrapado dentro de la cuerda de la elipse.

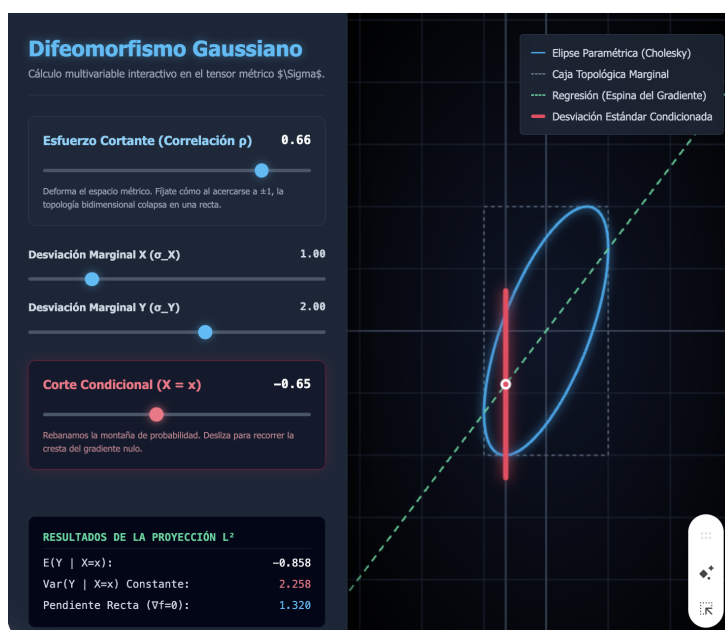
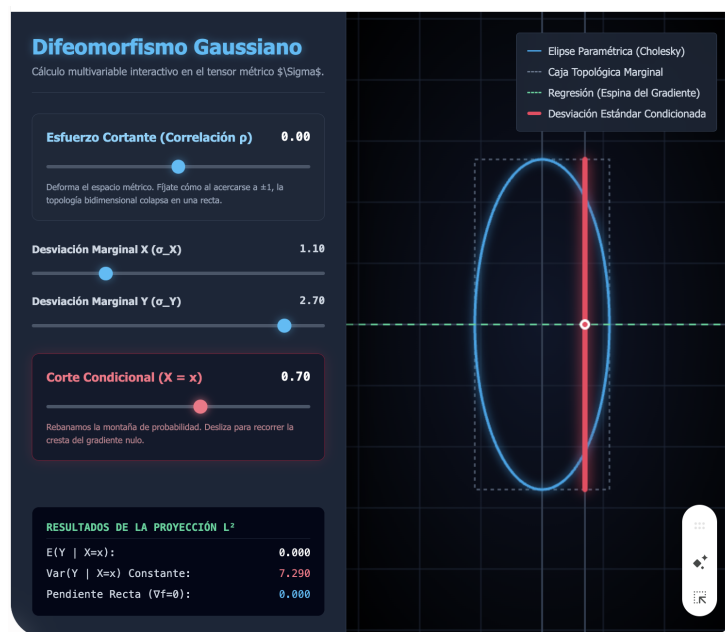
El Teorema de Pitágoras Topológico: La semilongitud de la cuerda horizontal en el centro exacto de una elipse inscrita en ese rectángulo obedece a la geometría pitagórica de las cónicas. Si el eje transversal máximo es 1 (que corresponde a σ_1^2), al inclinarla y estirla, una fracción de esa anchura se consume en mantener la diagonal. La fracción de anchura que se

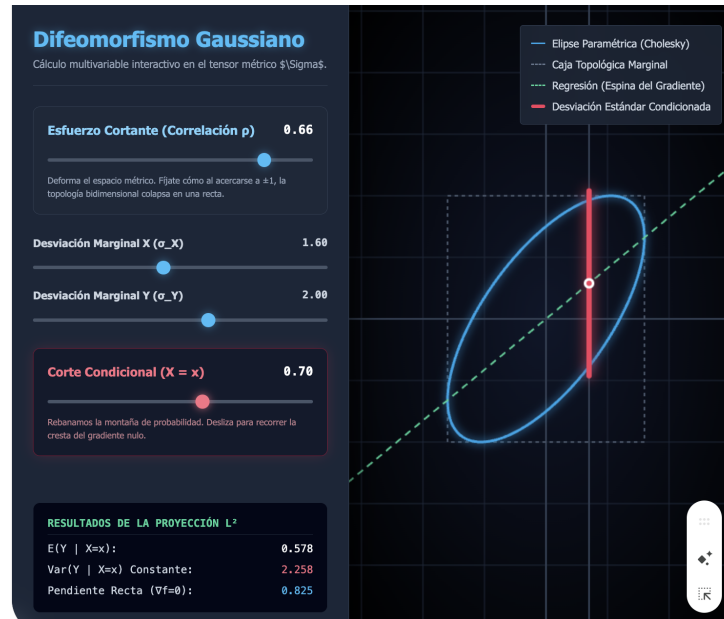
“come” la inclinación es exactamente ρ^2 .

El Remanente: Lo que queda libre para que te muevas horizontalmente de izquierda a derecha dentro de la cuerda es el total (1) menos lo que te ha bloqueado la inclinación (ρ^2). Por lo tanto, el ancho geométrico al cuadrado de esa porción rebanada es exactamente $\sigma_1^2(1 - \rho^2)$.

Ese $(1 - \rho^2)$ no es una casualidad algebraica; es el peaje geométrico que paga la elipse por inclinarse dentro de una caja fija. A medida que se inclina más (hacia $\rho = 1$), la elipse se vuelve tan fina y colapsa tanto sobre la diagonal que cualquier cuerda horizontal tiene longitud 0.

Esa cuerda horizontal que has cortado, con su centro desplazado proporcionalmente a la inclinación y su anchura estrangulada por esa misma inclinación, es la distribución normal condicionada.





3.3.2 El milagro lineal: la curva de regresión $h^*(x)$ en tres universos

1. El Milagro Lineal (Parte I): La Normal Bivariante

Ya hemos desentrañado este caso, pero veámoslo ahora bajo el prisma de h^* . La campana de Gauss no tiene “bordes”. Su probabilidad se desvanece hacia el infinito de forma perfectamente simétrica. Al rebanar la campana a cualquier altura, el centro de masa de esa rebanada se desliza armónicamente siguiendo la tracción de la covarianza. No hay paredes geométricas que empujen el centro de gravedad hacia un lado u otro. Por eso, en la Normal, $h^*(x)$ es estrictamente una recta, y la varianza (la anchura de la rebanada) es inmutable en todo el espacio.

2. El Milagro Lineal (Parte II): El Símplice Multinomial

La distribución Multinomial es radicalmente distinta de la Normal. Es discreta (vive en puntos aislados) y está enjaulada.

Imagina un sistema físico cerrado con 3 estados de energía y n partículas (por ejemplo, $n = 20$). Tienes X_1 , X_2 y X_3 , y una ley de conservación inviolable: $X_1 + X_2 + X_3 = n$. Si representamos esto en 2D (graficando X_1 frente a X_2 , puesto que X_3 es simplemente el resto), el espacio muestral es un triángulo de puntos delimitado por la frontera $x_1 + x_2 \leq n$. Es un símplice afín.

¿Qué pasa si condicionas X_2 a un valor fijo $X_1 = x$?

Le acabas de robar x partículas al universo. Al resto del sistema le quedan exactamente $(n - x)$ partículas para repartirse. La distribución condicionada $X_2 | X_1 = x$ resulta ser una Binomial pura con $(n - x)$ ensayos.

La esperanza de una Binomial es su número de ensayos por su probabilidad de éxito. Por tanto, el centro de gravedad es:

$$h^*(x) = \mathbb{E}(X_2 | X_1 = x) = (n - x) \frac{p_2}{1 - p_1}$$

¡Míralo bien! Matemáticamente es una función estrictamente lineal con pendiente negativa: $h^*(x) = \text{Constante} - (\text{Pendiente}) \cdot x$.

La Intuición Geométrica Brutal: En la Normal, la regresión era una recta por la simetría del infinito. En la Multinomial, la regresión es una recta por la geometría rígida de una pared. El “techo” del universo cae diagonalmente ($x_1 + x_2 = n$). A medida que avanzas en el eje X_1 , el espacio físico que le queda a X_2 se reduce de forma lineal. El centro de masa no tiene más remedio que aplastarse hacia el suelo siguiendo esa rampa triangular.

Además, fíjate en la varianza: $\text{Var}(X_2|X_1 = x) = (n - x)p(1 - p)$. Al revés que en la normal (donde la incertidumbre condicionada era constante), aquí la varianza condicionada colapsa hacia cero linealmente a medida que x se acerca a n . Acorralar el espacio aniquila el caos cuántico.



3. La Verdad No Lineal: La Bóveda Parabólica

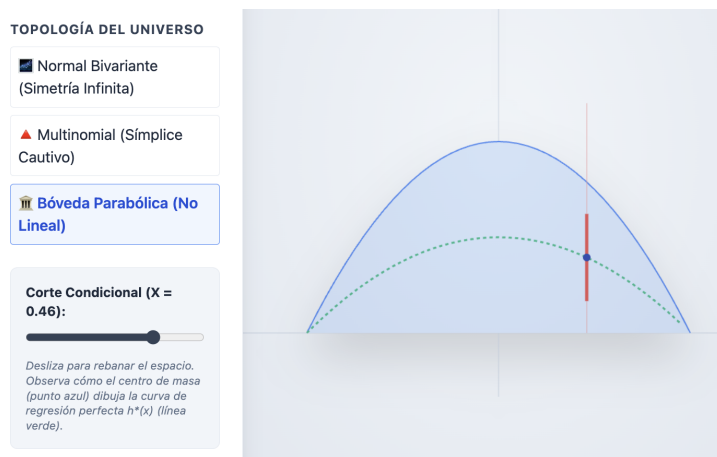
Para ver un caso donde $h^*(x)$ no es una recta, necesitamos romper la simetría afín y la linealidad de las paredes. Imagina una variable bivalente puramente geométrica: la probabilidad se distribuye uniformemente bajo una bóveda parabólica definida por $0 \leq Y \leq 1 - X^2$.

Si condicionas a un $X = x$ específico (un corte vertical), lo que te queda es una rebanada de materia continua (una distribución uniforme) que va desde el suelo ($Y = 0$) hasta el techo de la bóveda ($Y = 1 - x^2$).

El centro de masa (la esperanza condicionada) de una varilla uniforme está exactamente en su mitad. Por lo tanto:

$$h^*(x) = \mathbb{E}(Y|X = x) = \frac{1 - x^2}{2}$$

Aquí, $h^*(x)$ es una curva perfecta (una parábola invertida). La recta de mínimos cuadrados intentará pasar por el medio haciendo un promedio torpe, pero errará miserablemente. La verdadera conexión informacional entre Y y X está dictada por la curvatura del “techo” topológico del universo.



3.4 ¿Por qué la multinomial y la normal multivariante?

«Vale, vamos a ver la relación entre la multinomial y la normal multivariante. Sé perfectamente lo del símplex de la multinomial. Para mí, la multinomial (que modela la probabilidad de que conjuntamente tengas tantos elementos en tantas clases dado un conjunto de elementos y k clases) y la normal multivariante (que entiendo que generaliza la idea de una distribución centrada en una cierta media que condense bien cómo la incertidumbre suele distribuirse cuando el ruido es debido a muchos factores en varias dimensiones) modelan fenómenos tan diferentes que no tiene mucho sentido compararlas, ¿aun cuando efectivamente si N tiende a infinito en la multinomial el símplex se extiende y tiende a ser una normal multivariante de $N - 1$ dimensiones? ¿O hay más conexiones?

Por otro lado, en distribuciones unidimensionales he dado muchas, al menos 10. Sin embargo, en el caso multidimensional solo la multinomial y la normal multivariante. ¿Es esto un procedimiento estándar, o simplemente por falta de tiempo, o porque estas dos distribuciones condensan muchas cosas (quizás más que en 1D), o es que las otras no son muy generalizables, o qué?»

Imagina la función de masa de la multinomial como una “montaña” de probabilidad sobre una malla discreta.

Si tomamos el logaritmo de esta probabilidad (lo que en física estadística sería la entropía o la energía libre), obtenemos una superficie. En la cima de esa montaña (la moda, el estado de máxima probabilidad), si hacemos un Desarrollo de Taylor multivariante de segundo orden (cálculo diferencial de varias variables), la primera derivada se anula. ¿Qué nos queda? Una forma cuadrática pura dictada por la matriz Hessiana.

Al exponenciar una forma cuadrática pura (un paraboloide invertido), siempre obtienes el núcleo de una distribución normal multivariante: $\exp\{-\frac{1}{2}(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu})\}$.

Por lo tanto, la conexión no es solo que “tienden” a parecerse. Es que, topológicamente, cualquier variedad diferenciable parece euclidiana (plana) y cuadrática si la miras lo suficientemente cerca de su pico. La Normal Multivariante es, matemáticamente, el plano tangente estocástico a la variedad hiperbólica de la Multinomial.

En la Multinomial, la curva de regresión es: $E(X_i|X_j) = (n - X_j)\frac{p_i}{1-p_j}$. Es estrictamente lineal.

En la Normal Multivariante, la curva de regresión bidimensional es: $E(X_1|X_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)$. También es estrictamente lineal.

1. Tratabilidad Analítica (El milagro del álgebra lineal)

En 1D, puedes integrar y derivar casi cualquier distribución. En N dimensiones, el cálculo diferencial e integral se vuelve un infierno intratable. Sin embargo, estas dos distribuciones son matemáticamente “indestructibles”:

- **Marginalización perfecta:** Si ignoras variables en una normal multivariante, lo que queda sigue siendo normal. Si ignoras categorías en una multinomial, lo que queda es binomial.
- **Condicionamiento perfecto:** Si cortas la distribución normal fijando variables, la rebanada es otra normal. Si fijas una cuenta en la multinomial, la distribución condicional restante es binomial.
- **Transformaciones afines:** Si aplicas una transformación lineal a un vector normal multivariante ($\underline{X} = \underline{\mu} + A\underline{Z}$), el resultado sigue siendo normal multivariante.

Si intentas hacer esto con una distribución Gamma multivariante o una Beta multivariante (que sí existen, como la distribución de Dirichlet), el álgebra se rompe. Las integrales no tienen soluciones cerradas. Se enseñan estas dos porque son el “espacio vectorial base” donde las operaciones lineales funcionan limpiamente.

2. Teorema de Sklar y las Cópulas (La razón topológica real)

Esta es la respuesta filosófica profunda a tu pregunta de si «*las otras no son muy generalizables*».

En 1D modelamos fenómenos concretos: tiempos de espera (exponencial), fallos mecánicos (Weibull), eventos raros (Poisson).

En múltiples dimensiones, el problema matemático se divide en dos partes radicalmente distintas:

- **Las marginales:** Cómo se comporta cada variable sola.
- **La estructura de dependencia:** Cómo se mueven juntas.

El Teorema de Sklar (un pilar del análisis multivariante avanzado) demuestra que cualquier distribución multivariante se puede separar en sus marginales 1D y una función de dependencia pura llamada Cópula.

Resulta que la estructura de dependencia de la Normal Multivariante (la Cópula Gaussiana) es tan inmensamente flexible que, en la práctica profesional, tomamos distribuciones 1D extrañas (como Weibull o Gamma) y las “pegamos” usando la matriz de covarianzas de una Normal Multivariante. Se enseña la normal porque su “pegamento” (su matriz de covarianzas, definida como $\Sigma = E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^t$) es el estándar universal para modelar interacciones.

3. Atractores del caos (El Teorema Central del Límite Generalizado)

Acumular muchas variables extrañas en la naturaleza no genera “distribuciones Gamma multivariantes”. Sumar millones de procesos vectoriales aleatorios independientes (ya sean térmicos, biológicos o financieros) convergerá inexorablemente a la Normal Multivariante. Es un atractor topológico. La Multinomial es la base exacta del conteo; la Normal Multivariante es el destino final de la entropía.

Por eso, en un curso de Teoría de la Probabilidad como el tuyo, dominar estos dos pilares te da las llaves maestras para entender el 95% de la física estadística, el machine learning y la econometría espacial.

«Has pasado de algo discreto a un logaritmo continuo y lo has diferenciado, pero ¿qué es exactamente eso? Has cogido el logaritmo como podrías haber cogido, yo qué sé, un seno; ha sido arbitrario. En todo caso, en general, la historia suele ser que una distribución, siempre que la acotas en su segundo orden de las varianzas, se aproxima por una normal (que es pura media más varianza). De manera que lo que has dicho es solo un truco, es válido siempre y para toda distribución, no es una particularidad de la multinomial, ¿cierto?»

I. Por qué el logaritmo no es un seno: El isomorfismo de la naturaleza

Te preguntabas por qué usar el logaritmo y no un seno u otra función cualquiera. En matemáticas avanzadas y en termodinámica, el logaritmo jamás es una elección arbitraria; es el único puente estructural entre dos mundos que hablan idiomas distintos.

- **El mundo de la Probabilidad (Multiplicativo):** La probabilidad de que ocurran eventos independientes se rige por la multiplicación. Si tiras dos dados, la probabilidad conjunta es $P(A \cap B) = P(A) \cdot P(B)$.
- **El mundo de la Física y la Geometría (Aditivo):** Sin embargo, la geometría y la física clásica se construyen sumando. Las distancias en un mapa se suman. Las energías en un gas se suman. Los vectores en un plano se suman.

Si queremos usar las herramientas de la geometría diferencial (curvaturas, planos tangentes, derivadas espaciales) en la probabilidad, necesitamos un “traductor” matemático que convierta multiplicaciones en sumas de manera perfecta. El logaritmo es el único isomorfismo que hace esta magia: $\log(a \cdot b) = \log(a) + \log(b)$.

Si hubiéramos tomado un seno, obtendríamos $\sin(a \cdot b)$, que geométricamente es ruido blanco sin estructura. Al tomar el logaritmo de una función de masa de probabilidad, estamos transformando “probabilidades apiladas” en “energías potenciales”. De ahí nace la Entropía de Shannon o la mecánica estadística de Boltzmann ($S = -k \sum p \log p$). Transformamos el tejido discreto de las probabilidades en un paisaje de colinas y valles de “información” donde ya podemos usar geometría real.

II. De la malla discreta a la sábana continua: El límite termodinámico

Me pides que aclare ese salto brusco de lo discreto a lo continuo.

Visualiza la distribución Multinomial (que cuenta combinaciones de bolitas cayendo en k cajas) como una red de pesca hecha de nudos discretos de cuerda. Cada nudo es una combinación posible exacta. No puedes tener 3.5 bolitas en la caja A; estás atrapado en las intersecciones de la red.

A medida que el número de bolitas (N) crece hacia el infinito (lo que en física llamamos el límite termodinámico), esos nudos se aprietan más y más, multiplicándose exponencialmente. Si te alejas visualmente, la red de pesca deja de parecer una malla de hilos discretos y se funde en una tela visualmente continua, una sábana elástica sobre la que podemos trazar tangentes.

El puente matemático que permite derivar esta sábana se llama Aproximación de Stirling. Sabemos que la Multinomial está llena de factoriales (que son puramente discretos). Stirling demostró que en el infinito, los factoriales se comportan como continuos a través del logaritmo: $\ln(N!) \approx N \ln(N) - N$. Al aplicar este “logaritmo no arbitrario” a la Multinomial y hacer crecer N , los nudos discretos se licúan algebraicamente, desaparecen los saltos, y emergen polinomios cuadráticos continuos. La red se hace sábana.

III. Tu gran objeción: ¿Es solo el truco general de la varianza?

Afirmas que, en general, siempre que acotas cualquier distribución en su segundo orden (derivada segunda, que modela la varianza), se aproxima por una normal. Dices que es válido siempre y no una particularidad de la Multinomial.

Tienes toda la jodida razón. Matemáticamente, esto es el Teorema Central del Límite (o la aproximación asintótica de Laplace). Pero aquí es donde entra la Geometría de la Información (Information Geometry), una rama bestial que une la estadística con la Relatividad General de Einstein, para darte la intuición definitiva de por qué esto ocurre.

Imagina que el espacio de todas las distribuciones de probabilidad posibles es un universo curvo (una variedad de Riemann hiperbólica).

Cada familia (las de Poisson, las Binomiales, las Exponenciales) vive en una latitud diferente de este espacio curvo.

¿Cómo medimos las “curvas” en este espacio? Usamos el tensor métrico, que aquí se llama la Métrica de Información de Fisher. Esta matriz geométrica está formada, exactamente, por las segundas derivadas del logaritmo de la probabilidad.

Hacer un desarrollo de Taylor de segundo orden en el pico de una probabilidad significa, geoméricamente, plantar un plano tangente euclidiano (plano y recto) apoyado sobre esa superficie curva. Como la Tierra es redonda, pero a escala local (en el jardín de tu casa) te parece matemáticamente plana, cualquier variedad estadística “curva” se aproxima por un espacio “plano” a nivel microscópico si miras solo su pico.

¿Y cuál es la única distribución matemática que existe naturalmente en un espacio plano y euclidiano sin curvaturas raras? La Normal.

Por lo tanto, sí, la Normal Multivariante es el “plano tangente genérico” de casi cualquier distribución que se comporte bien.

3.5 Ejemplos

Preguntabas por un caso representativo multivariante de mayor orden que sea verdaderamente útil. La respuesta es el Santo Grial de la modelización espacial y financiera: Los Procesos Gaussianos y la Econometría de Carteras (ej. $N = 500$ dimensiones).

Imagina las 500 empresas del índice S&P 500 del mercado bursátil. Tienes un vector aleatorio gigantesco: $\underline{X} = (X_1, X_2, \dots, X_{500}) \sim \mathcal{N}_{500}(\underline{\mu}, \Sigma)$.

La matriz de covarianzas Σ es una bestia de 500×500 .

¿Cómo se lee esto visualmente? Es un hiperelipsoide en un espacio de 500 dimensiones.

La estructura oculta: En la realidad, las acciones de empresas tecnológicas (Apple, Microsoft, Google) tienen mucha ρ entre ellas. Los bancos tienen mucha ρ entre ellos. Pero una tecnológica y un campo de trigo en Kansas tienen poca ρ .

La reducción topológica: Aunque vives en 500 dimensiones, la matriz de covarianzas suele estar “aplastada” en la mayoría de sus direcciones. Las primeras “direcciones principales” (autovectores) explican casi todo. El primer autovector suele ser el “Riesgo Global del Mercado” (el día que hay pánico mundial, todo colapsa, ρ generalizada hacia 1). El hiperelipsoide es en realidad una figura casi plana que vive en apenas 5 o 6 dimensiones útiles dentro de un abismo de 500 dimensiones vacías.

La Normal Multivariante es tan colosalmente útil porque, al permitir que todo interactúe con todo de forma lineal, la matriz de covarianzas Σ actúa como un mapa topológico. Las variables altamente correlacionadas están matemáticamente “cerca”, y las incorrelacionadas son “ortogonales” (perpendiculares).

«No, eso ya lo sé. Quiero que, en cambio, pongas un ejemplo (por ejemplo, trivariante) del mundo científico donde exista una cierta correlación entre dos variables. ¿Y qué significa intuitivamente (no geoméricamente, eso ya lo sé) diagonalizar o encontrar las variables aleatorias en la dirección de los autovectores de Σ ? Queriendo decir, no sé, por ejemplo, si yo tengo altura, edad y probabilidad de morir y veo la correlación, ¿qué quiere decir diagonalizar? ¿Mis nuevas X_i pueden tener un significado muy claro, o es una ponderación de las tres que a nosotros no nos resulta intuitiva? ¿O hay casos en que sí?»

Has tocado la fibra sensible de la filosofía de la ciencia. Lo que estás preguntando es, literalmente, cómo los científicos descubren conceptos que no pueden ver ni tocar. Diagonalizar una matriz de covarianzas Σ no es solo un truco de álgebra lineal o una rotación de ejes; es el proceso de encontrar el código fuente de la realidad.

I. El Problema Trivariante Real: Los Fotorreceptores de tu Ojo

En el fondo de tu retina tienes tres tipos de células fotorreceptoras (conos) que reaccionan a la luz. Este es nuestro vector aleatorio trivariante natural, $\underline{X} = (X_1, X_2, X_3)^t$:

- X_1 : Conos L (Sensibles a longitudes de onda largas / Rojos).
- X_2 : Conos M (Sensibles a longitudes medias / Verdes).
- X_3 : Conos S (Sensibles a longitudes cortas / Azules).

Imagina que miras un bosque o una calle. Llega luz a tu ojo y genera tres señales.

El problema físico: En la naturaleza, la luz rara vez excita un solo cono. Los espectros de absorción de los conos L y M están absurdamente solapados. Biológicamente, si se activa el cono L (rojo), el cono M (verde) casi siempre se activa también.

Si construimos la matriz de covarianzas empírica Σ de la luz que entra al ojo humano en un entorno natural, vemos una correlación masiva: $\rho_{LM} \approx 0.90$.

El nervio óptico es un “cable” con un ancho de banda limitado. Si el ojo enviara al cerebro el vector original (L, M, S) , estaría enviando información redundante todo el tiempo. Estaría desperdiciando energía.

II. La Solución de la Naturaleza: Diagonalizar Σ

En 1983, los investigadores Buchsbaum y Gottschalk publicaron un artículo histórico (*Derivation of a color space from natural scenes*). Demostraron matemáticamente que las células ganglionares de nuestra retina no envían la señal bruta (L, M, S) al cerebro. Lo que hacen es

diagonalizar la matriz de covarianzas de la luz natural (lo que en estadística matemática se llama Análisis de Componentes Principales o PCA).

El cerebro calcula los autovectores de esa matriz Σ para encontrar tres nuevas variables aleatorias (Z_1, Z_2, Z_3) que sean matemáticamente independientes y ortogonales (correlación cero).

¿Qué significan intuitivamente estos autovectores matemáticos en tu vida diaria? Tienen un significado tan real que definen la realidad humana:

1. Primer Autovector (Z_1): La suma ponderada (Luminancia)

El primer autovector que extraen las matemáticas tiene coeficientes positivos para L y M. La combinación lineal es aproximadamente:

$$Z_1 = L + M$$

Su significado intuitivo: Las matemáticas se dan cuenta de que, dado que L y M se mueven juntos casi siempre, la dirección de mayor varianza en el mundo es simplemente “la cantidad total de luz”. Este autovector es el Canal de Luminancia (Blanco y Negro).

Es la imagen en escala de grises. Fíjate en lo bestial que es esto: la “Luz” o el “Brillo” no existe como un receptor físico en tu ojo (solo tienes detectores de color); el Brillo es un concepto matemático puro, es el primer autovector de la matriz de covarianzas del mundo visual.

2. Segundo Autovector (Z_2): La resta ortogonal (Oponente Rojo-Verde)

El segundo autovector, por obligación del álgebra lineal, tiene que ser perpendicular al primero. Las matemáticas buscan la siguiente dirección de mayor varianza que sea independiente del brillo. Los coeficientes resultan ser de signos opuestos:

$$Z_2 = L - M$$

Su significado intuitivo: Este es el Canal Oponente Rojo-Verde. Las matemáticas restan la señal verde de la roja.

Intuitivamente, esto responde a una pregunta milenaria de la psicología: ¿Por qué puedes imaginar un “verde azulado” o un “amarillo anaranjado”, pero es biológicamente imposible imaginar un “verde rojizo”?

Porque el Rojo y el Verde no son variables independientes en tu cerebro. Son los dos extremos del mismo autovector matemático (Z_2). Tu cerebro solo puede recibir un valor numérico para Z_2 : si es positivo, ves rojo; si es negativo, ves verde. No puedes estar en los dos sitios del autovector a la vez.

3. Tercer Autovector (Z_3): El residuo independiente (Oponente Azul-Amarillo)

El tercer autovector es ortogonal a los dos anteriores. Su combinación lineal agrupa a L y M y los enfrenta a S:

$$Z_3 = S - (L + M)$$

Su significado intuitivo: Este es el Canal Oponente Azul-Amarillo. (Notemos que L+M es lo que percibimos como amarillo). Las matemáticas han aislado la única porción de varianza de la luz que el canal rojo/verde y el blanco/negro no lograban explicar.