Memorando completo: distribución normal, densidad, σ, varian

Recopilación completa de la conversación: desde la intuición de la campana hasta cálculos de varianza/desviación, diferencias entre densidad y probabilidad, visualización en Python/Calc, y por qué la normal aparece tanto (TCL).

Objetivo y contexto

Estabas aprendiendo estadística básica para aplicarla en Python (fintech/inversión). Un vídeo corto y denso no te encajaba; aquí fuimos paso a paso con ejemplos, código y visualización hasta que las ideas clave quedaron claras.

1) La campana: qué es la distribución normal

La normal (o gaussiana) es una distribución continua con forma de campana. La mayoría de valores caen cerca del centro y los extremos son cada vez menos frecuentes. Se usa mucho como modelo y como aproximación.

2) Parámetros μ y σ (y qué controla cada uno)

• μ (mu) es la media: el centro.
• σ (sigma) es la desviación típica: la escala de dispersión (anchura).

Efecto visual:
• Cambiar μ desplaza la campana.
• σ pequeño ⇒ campana alta y estrecha (más concentración).
• σ grande ⇒ campana baja y ancha (más dispersión).

El pico en x=μ vale 1/(σ√(2π)). Por eso, si σ sube, el pico baja: el área total debe seguir siendo 1.

3) Densidad (altura) vs probabilidad (área): lo que más bloquea al principio

La curva f(x) es una DENSIDAD, no una probabilidad.

En continuo:
• La probabilidad en un punto exacto es 0: P(X=x)=0.
• La probabilidad real es el área bajo la curva en un intervalo: P(a≤X≤b).

Aproximación para entenderlo:
Si tomas un intervalo pequeño Δx, entonces P(x..x+Δx) ≈ f(x)·Δx.

Por eso f(x) puede ser >1 (no es probabilidad), y la suma total de probabilidades sigue siendo 1 porque el área total bajo la curva es 1.

4) Fórmulas que usamos (sin meternos en “por qué” avanzando)

Estas expresiones son las que aparecieron repetidamente en código y razonamiento.

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

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2}$$

$$\Phi(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

5) Regla 68–95: “rango típico” y concentración

Para una normal:

• Entre $\mu-\sigma$ y $\mu+\sigma$ cae ≈ 68.27%.

• Entre $\mu-2\sigma$ y $\mu+2\sigma$ cae ≈ 95.45%.

Lectura correcta:

• En una normal, $\pm 1\sigma$ es un rango “típico” (convención útil).

• σ define la escala: si σ cambia, el intervalo en X cambia, pero la fracción (68% / 95%) se mantiene.

Importante: fijarse en ángulos tipo 45° en la pendiente NO es relevante estadísticamente: depende del zoom y de los ejes. Lo intrínseco es el concepto de “a cuántos sigmas del centro” estás.

$$P(\mu - \sigma \le X \le \mu + \sigma) \approx 0.6827$$

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 0.9545$$

6) σ negativo y σ=0

• σ negativo no tiene sentido: la densidad saldría negativa por el factor 1/σ (no sería una distribución válida).
• σ=0 tampoco es una normal continua: es el límite donde toda la probabilidad se concentra en μ (distribución degenerada).

7) Varianza y desviación típica (población total)

Primero hablamos de población total: tienes TODOS los datos, no estás estimando.

Definiciones:
• $\mu = (1/N)\ \Sigma\ x_i$
• $\sigma^2 = (1/N)\ \Sigma\ (x_i - \mu)^2$
• $\sigma = \sqrt{\sigma^2}$

Punto clave:
La desviación típica estándar SIEMPRE es la raíz de la varianza. Puedes escribir σ con sumatorio, pero siempre incluye esa raíz.

8) Por qué elegiste 3 (luego 4) poblaciones con la misma media

Elegiste conjuntos con μ=5 para aislar el efecto de la dispersión.

A: 1,3,5,7,9 → var=8, σ≈2.828
B: 1,2,5,8,9 → var=10, σ≈3.162 (más dispersa)
C: 1,4,5,7,8 → var=6, σ≈2.449 (más concentrada)
D: 3,4,5,6,7 → var=2, σ≈1.414 (muy concentrada)

Al ajustar una campana a cada una (usando μ y σ), se ve claro: mismo centro, distinta anchura/altura.

9) Probabilidad en un intervalo visual: P(3..7)

Tu intuición fue correcta: con μ=5 fijo, cuanto menor sea σ, mayor es la probabilidad de caer cerca del centro en un intervalo fijo como [3,7].

También diferenciamos:
• Probabilidad empírica (discreta): contar cuántos valores caen en [3,7]/N.
• Probabilidad del modelo normal (continua): área bajo la campana.

Con N=5 pueden parecerse a veces, pero no son lo mismo.

10) Teorema Central del Límite (TCL) y el script de dados

El TCL explica por qué aparece la normal: sumas (o promedios) de muchos efectos pequeños e independientes tienden a una normal.

Lo viste con dados:
• 1 dado: uniforme.
• 2–3: empieza a curvarse.
• 10–15: campana clara.

• 100: muy parecido a normal.

Esto conecta con ejemplos reales (altura humana) donde se suman genética + ambiente + hábitos + ruido.

11) Casos donde NO es simétrica (no normal)

Ingresos/riqueza, tiempos de espera, tamaños de archivos: suelen ser sesgados y con colas largas. Ahí la normal puede fallar incluso con muchos datos; suelen encajar mejor lognormal/gamma, mezclas, etc.

12) Implementación en Python: lo que aclaramos

• np.linspace(a,b,n) devuelve un array con n puntos equiespaciados.

• NumPy vectoriza: f = ... te da un array de resultados, no una función tipo JS.

• ax.plot devuelve lista de líneas; “line, = ...” es unpacking.

• ax.plot([],[]) crea la línea vacía para animar con line.set_data.

• FuncAnimation llama update(frame) por frame.

• Si haces ax.set_ylim en cada frame, cambia la escala del eje Y (la “altura” visual), no el área/probabilidad.

También añadimos líneas $\mu\pm\sigma$, $\mu\pm2\sigma$, rellenos $\pm1\sigma/\pm2\sigma$ y textos 68.3% / 95.4%.

13) LibreOffice Calc: hoja para barrer σ

Montaste una hoja para ver cómo cambia la campana al variar σ. Arreglamos el problema de que N/x_min/x_max no parecían actualizar ampliando el rango y usando NA() fuera de N para que el gráfico ignore lo que sobra.

Curiosidad: el valor donde numéricamente coincide “σ = altura del pico” sale de $1/(\sigma\sqrt{(2\pi)})=\sigma \Rightarrow \sigma\approx0.631$. No es “místico”; es una coincidencia de escala.

Cierre: lo que ya dominas y por qué te sirve en fintech

Ahora ya manejas los conceptos base: μ, σ, varianza; densidad vs área; y por qué $\pm1\sigma/\pm2\sigma$ son reglas prácticas. En finanzas esto conecta con volatilidad (σ), z-scores y detección de eventos raros. Ojo: retornos reales a menudo no son normales (colas más pesadas), pero entender la normal es el punto de partida correcto.