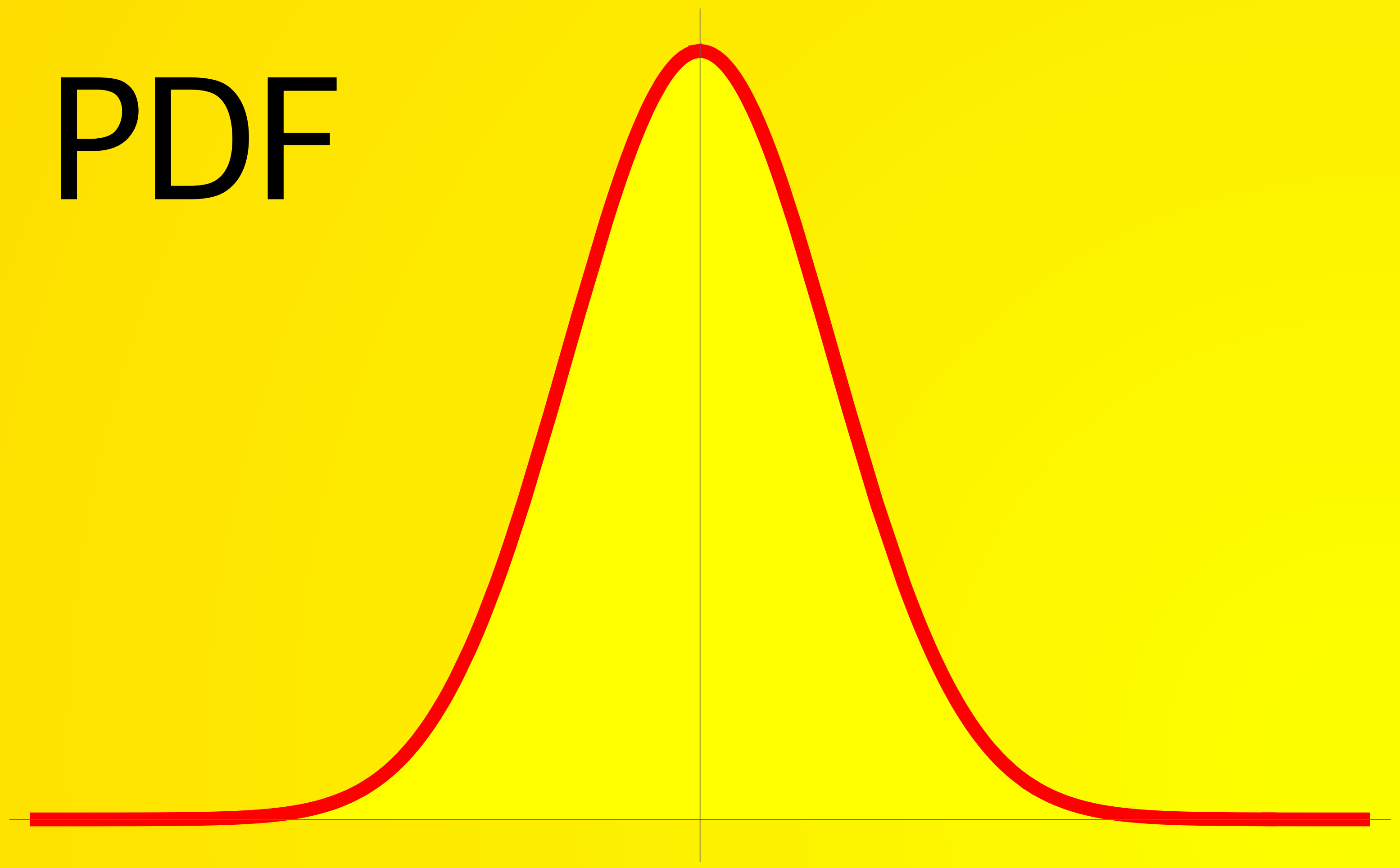
30

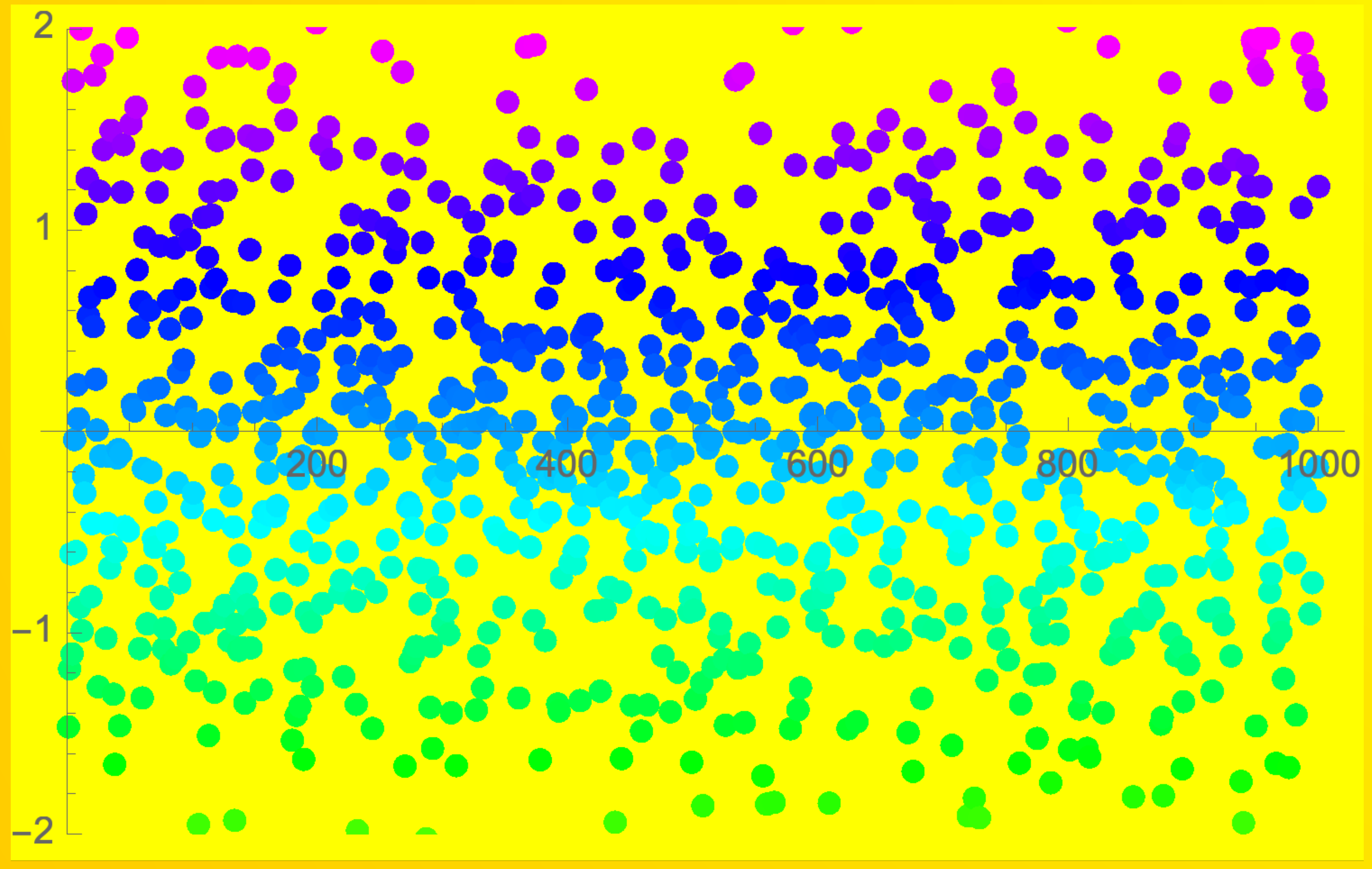
Statistics

PDF

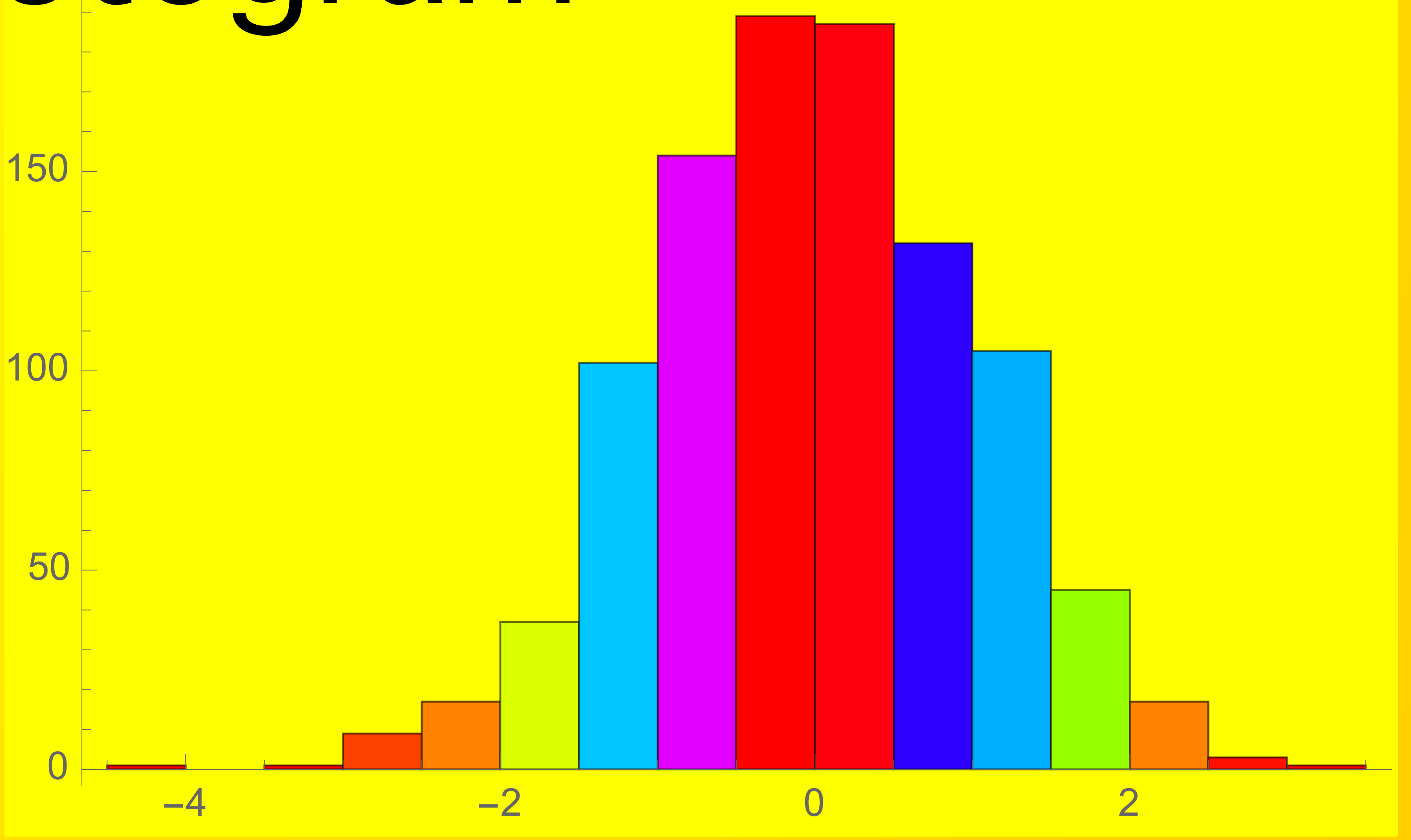


CDF

Data



Histogram



PLAN

1. Poll

2. PDF reminder

3. From data to PDF

4. Moments

5. Variance

POLL

Which one is the correct formula for the standard normal distribution?

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

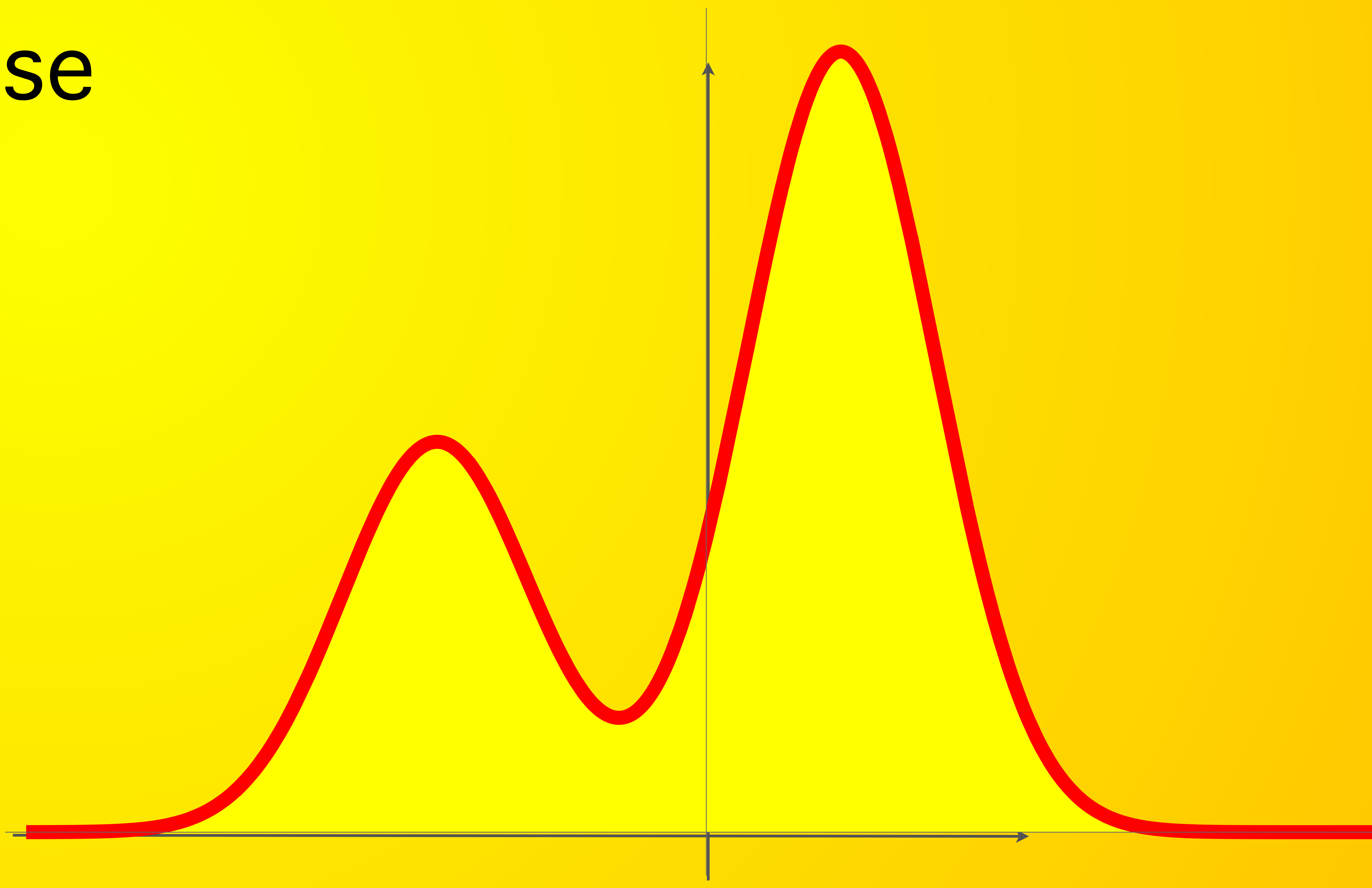
B $$f(x) = \frac{e^{-x^2/2}}{\sqrt{\pi}}$$

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

PDF Reminder

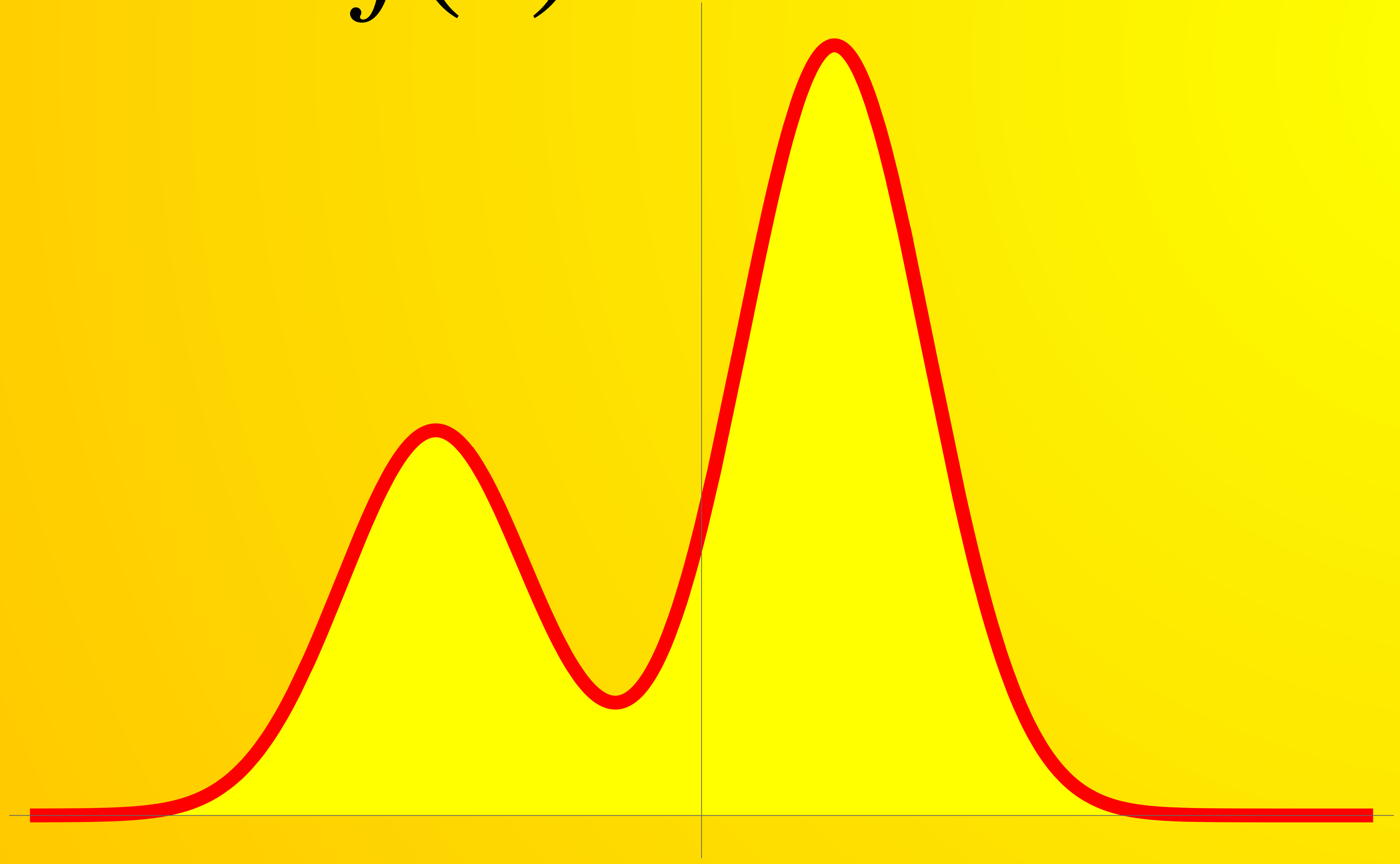
- The function is nonnegative
- The function is piecewise continuous
- The total integral is 1

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

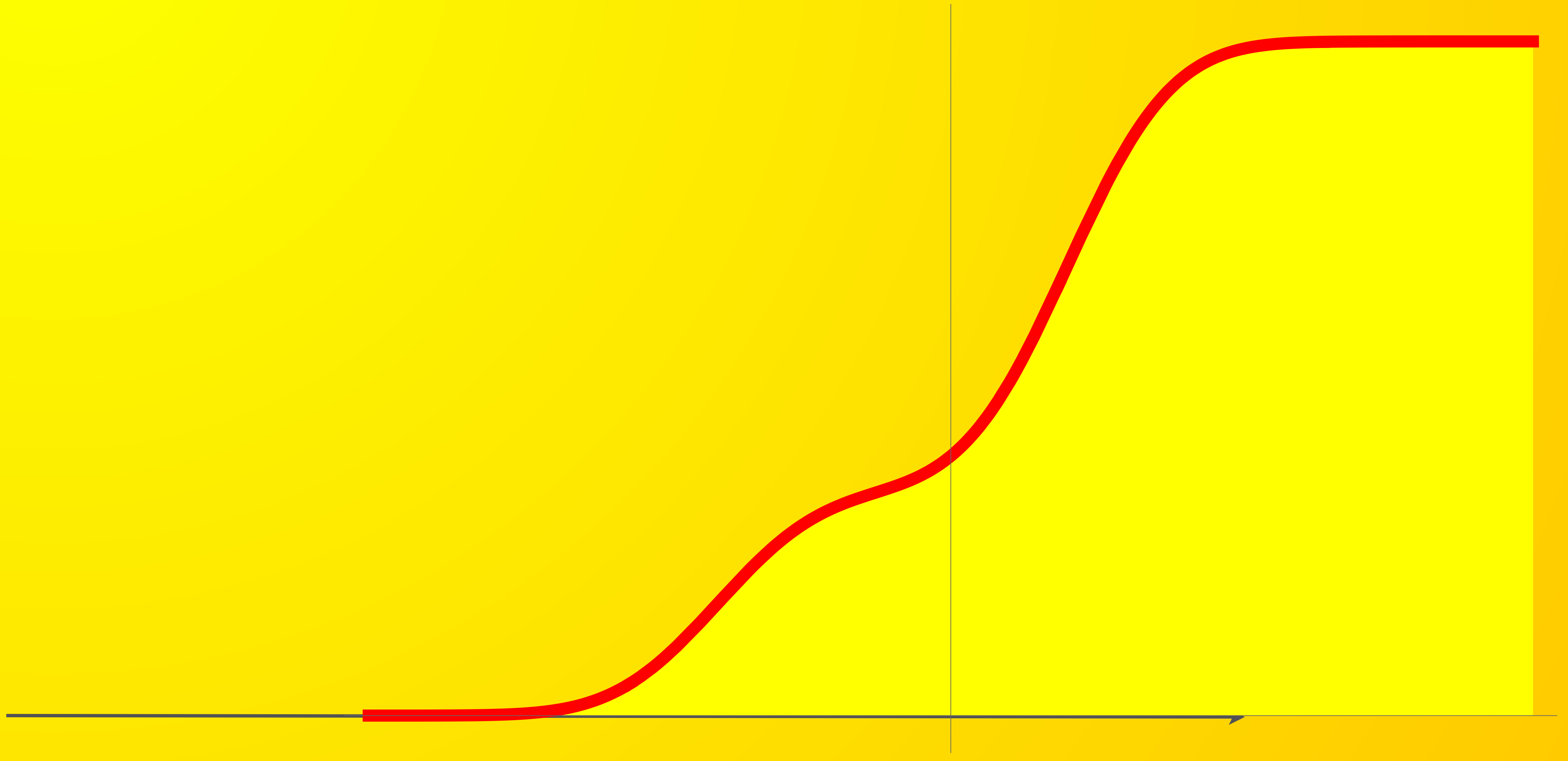


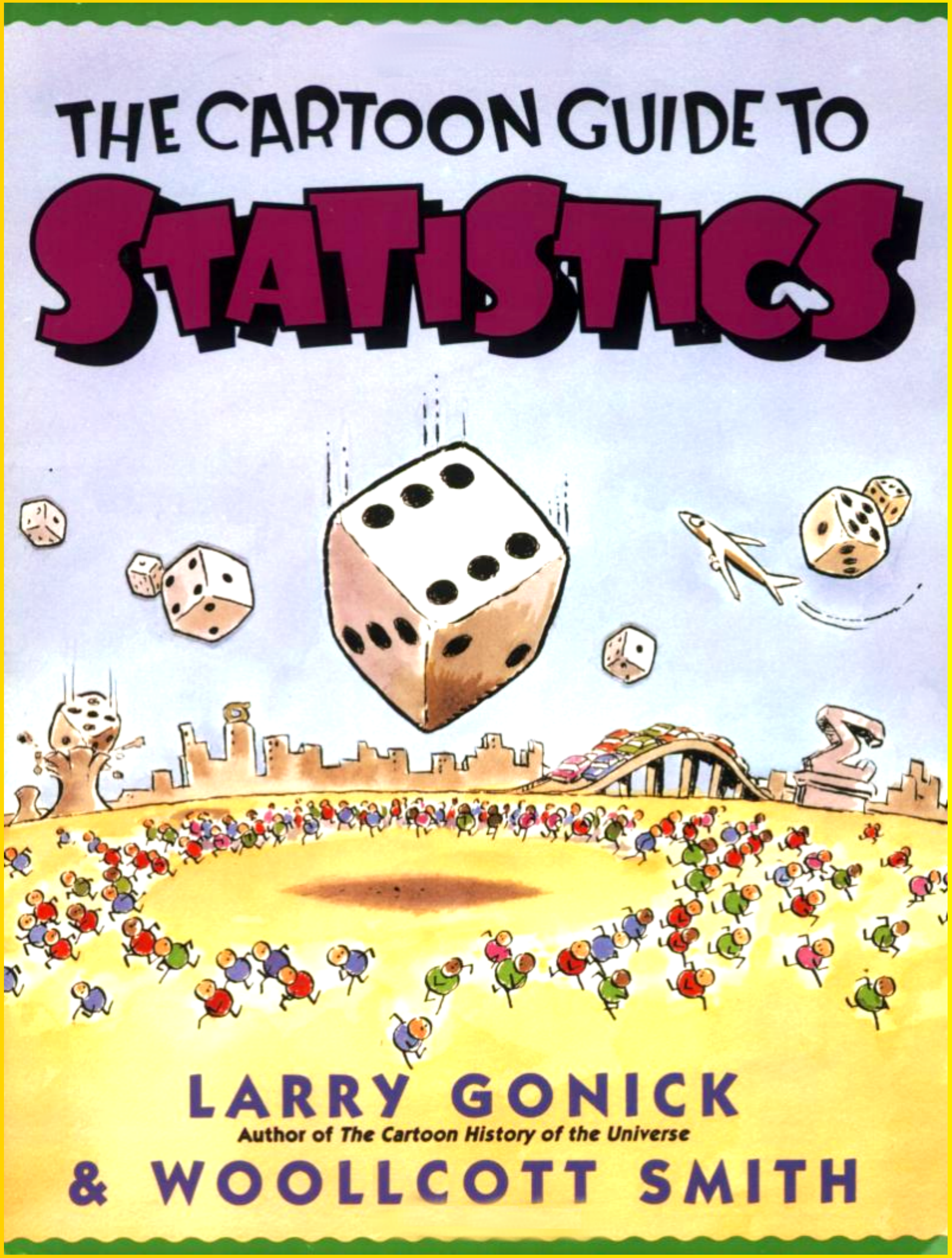
CDF Reminder

$f(x)$



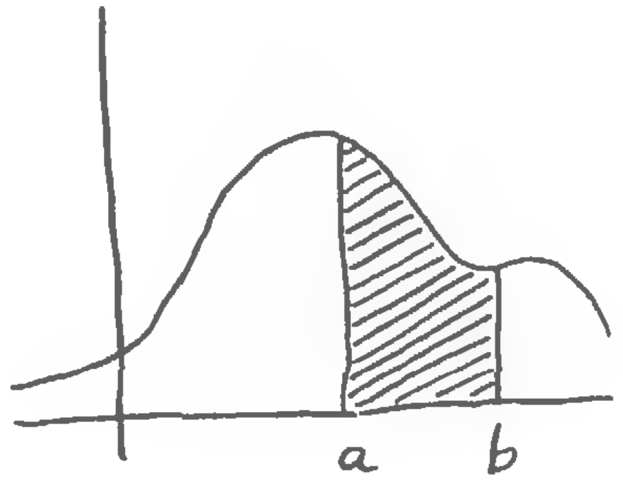
$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$





IN GENERAL, THE PROBABILITY DENSITY WON'T BE SO SIMPLE, AND COMPUTING THE AREAS CAN BE FAR FROM TRIVIAL.

$$\int_a^b f(x)\,dx$$



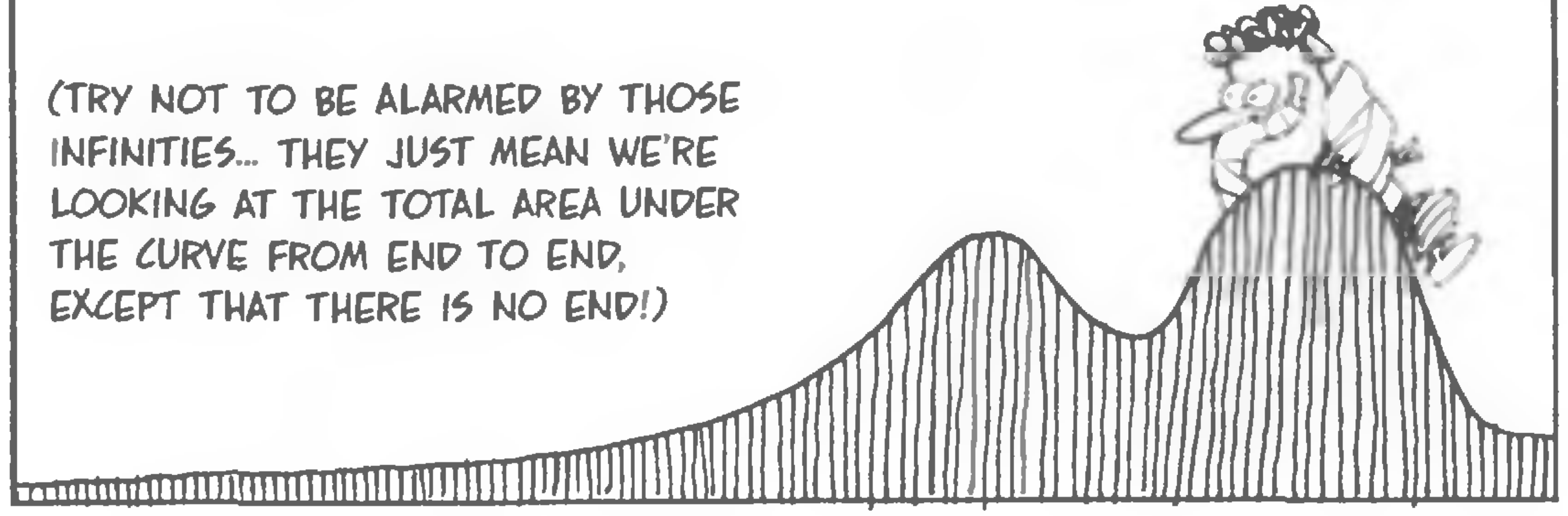
WE HAVE TO USE CALCULUS NOTATION TO DESCRIBE THE AREA UNDER THE CURVE $f(x)$. THIS SYMBOL IS READ "THE INTEGRAL OF f FROM a TO b."

LIKE DISCRETE PROBABILITIES, CONTINUOUS DENSITIES HAVE TWO FAMILIAR PROPERTIES:

$$f(x) \geq 0$$

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

(TRY NOT TO BE ALARMED BY THOSE INFINITIES... THEY JUST MEAN WE'RE LOOKING AT THE TOTAL AREA UNDER THE CURVE FROM END TO END, EXCEPT THAT THERE IS NO END!)

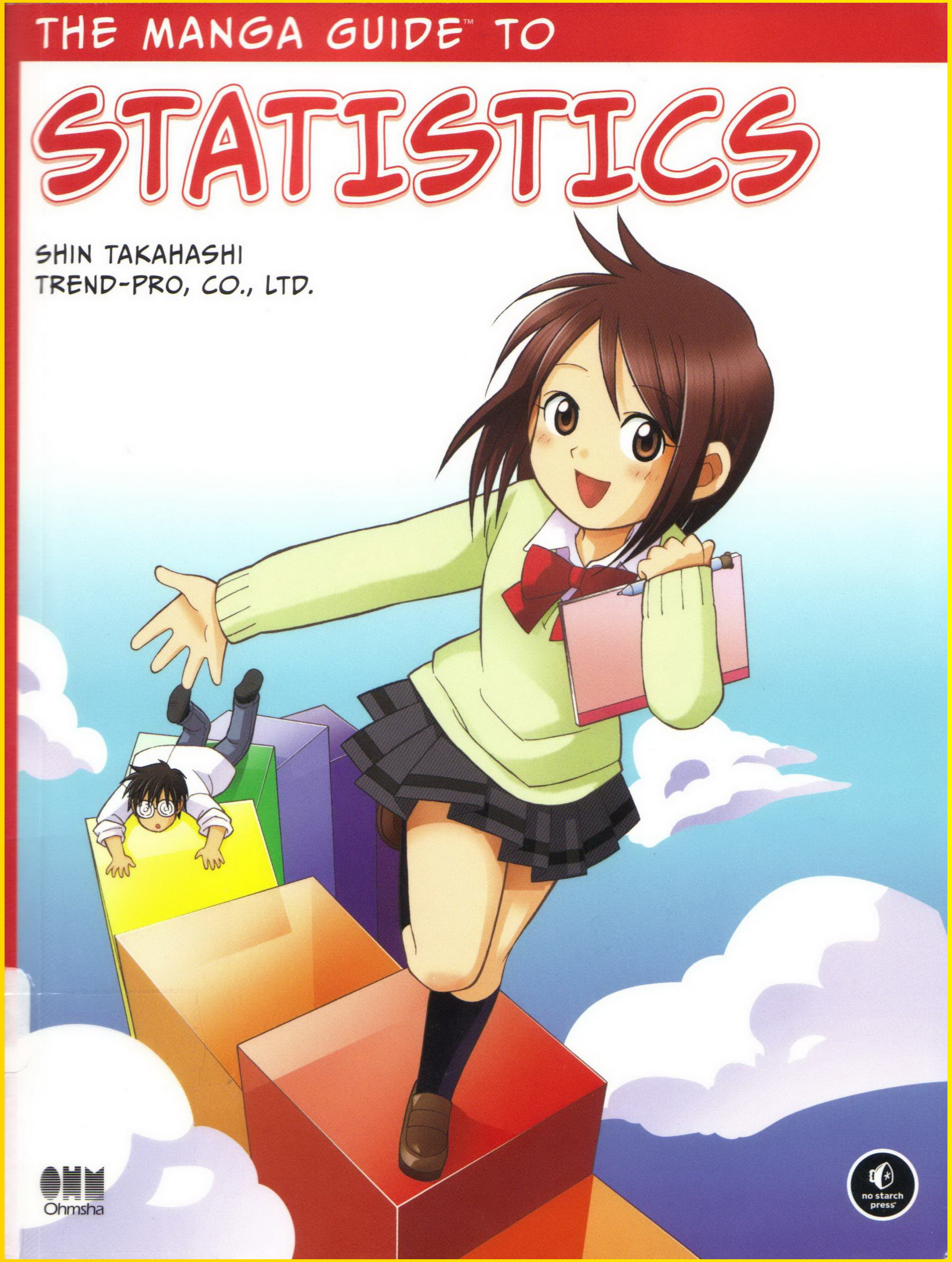


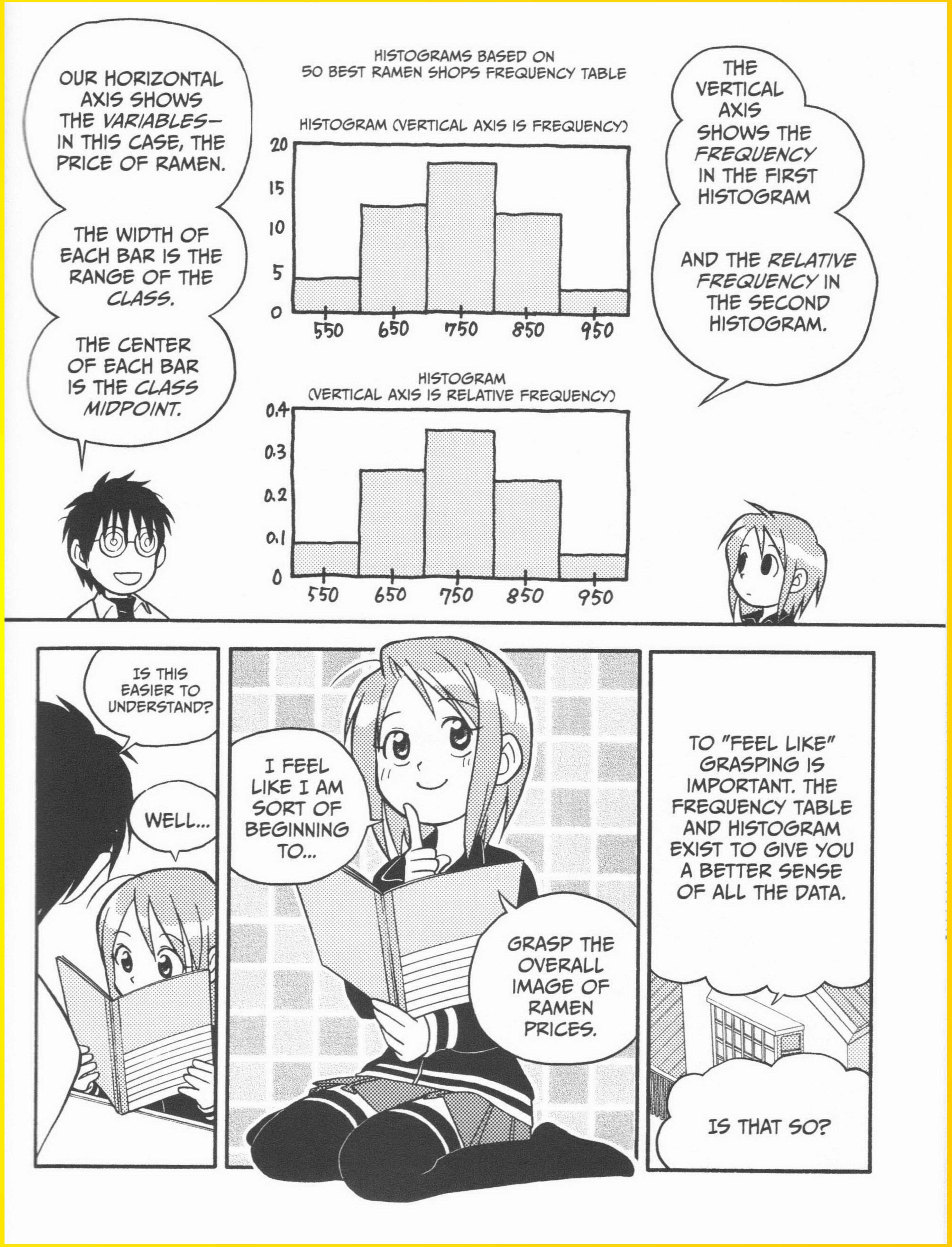
Data -> Functions

1000 data points
Data

smoothed PDF

Histogram

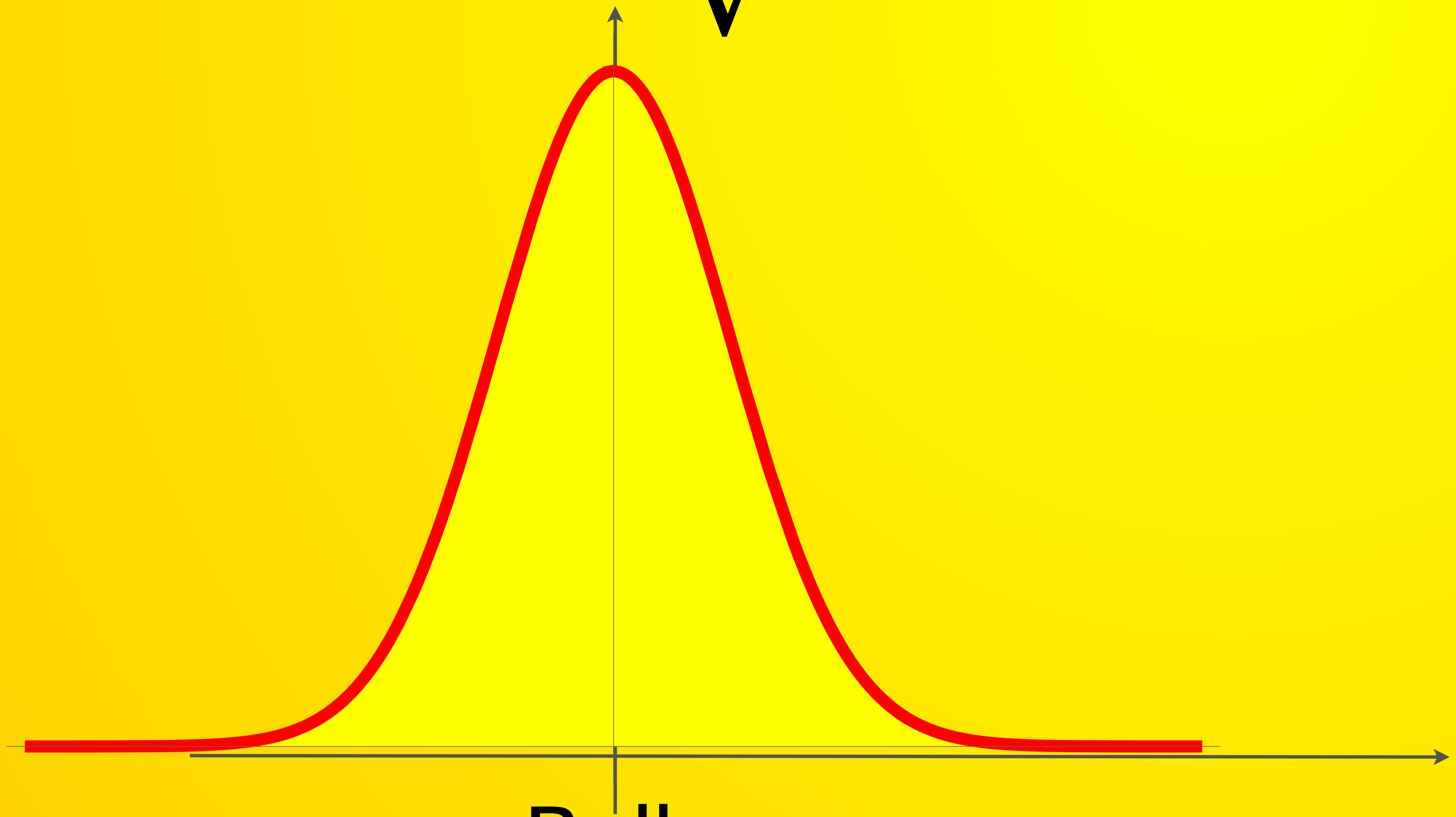

THE MANGA GUIDE™ TO
STATISTICS
SHIN TAKAHASHI
TREND-PRO, CO., LTD.
Ohmsha
no starch press


OUR HORIZONTAL AXIS SHOWS THE *VARIABLES*—IN THIS CASE, THE PRICE OF RAMEN.
THE WIDTH OF EACH BAR IS THE RANGE OF THE *CLASS*.
THE CENTER OF EACH BAR IS THE *CLASS MIDPOINT*.
HISTOGRAMS BASED ON 50 BEST RAMEN SHOPS FREQUENCY TABLE
HISTOGRAM (VERTICAL AXIS IS FREQUENCY)
20
15
10
5
0
550
650
750
850
950
HISTOGRAM (VERTICAL AXIS IS RELATIVE FREQUENCY)
0.4
0.3
0.2
0.1
0
550
650
750
850
950
THE VERTICAL AXIS SHOWS THE *FREQUENCY* IN THE FIRST HISTOGRAM
AND THE *RELATIVE FREQUENCY* IN THE SECOND HISTOGRAM.
IS THIS EASIER TO UNDERSTAND?
WELL...
I FEEL LIKE I AM SORT OF BEGINNING TO...
GRASP THE OVERALL IMAGE OF RAMEN PRICES.
TO "FEEL LIKE" GRASPING IS IMPORTANT. THE FREQUENCY TABLE AND HISTOGRAM EXIST TO GIVE YOU A BETTER SENSE OF ALL THE DATA.
IS THAT SO?

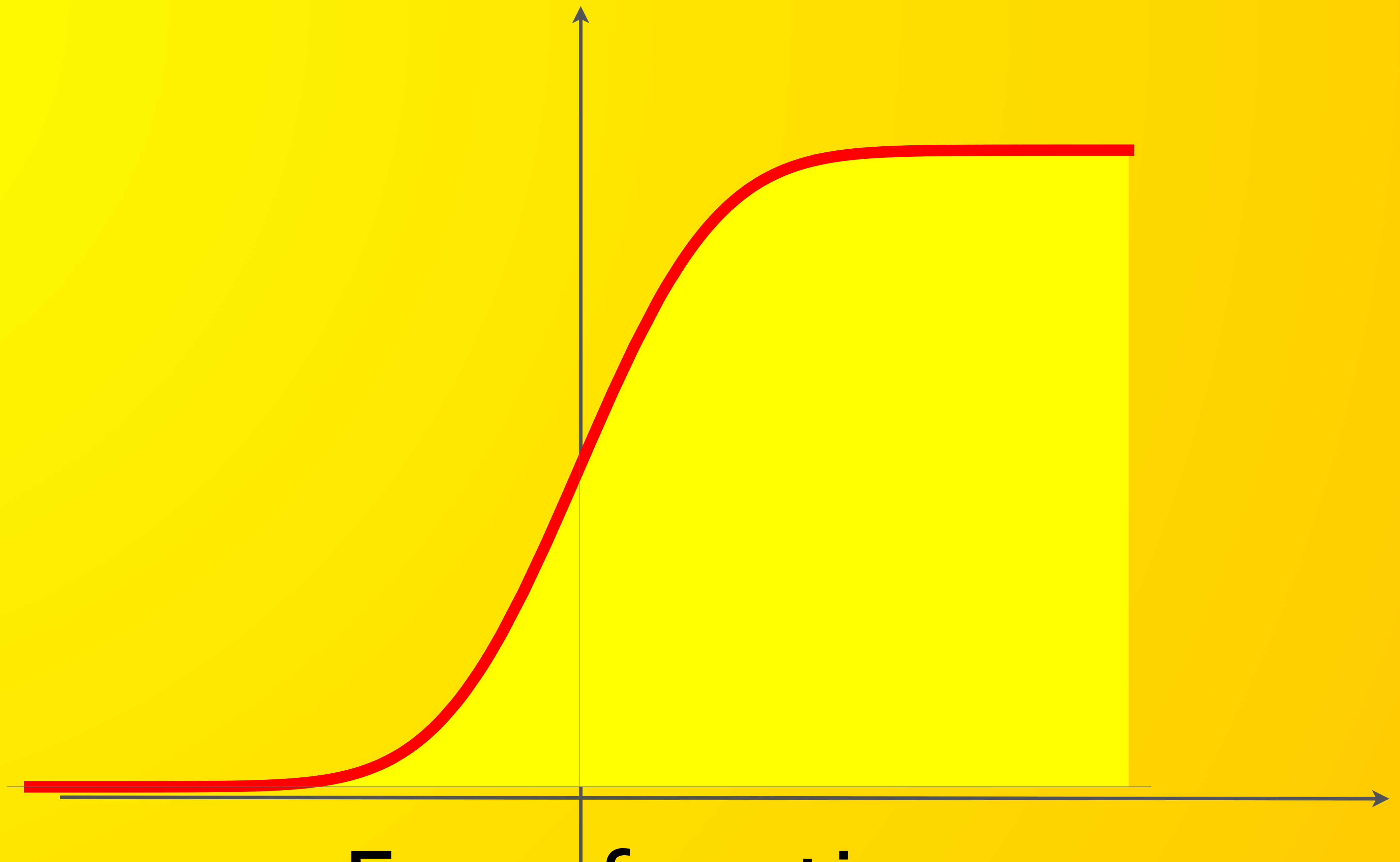
NORMAL

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

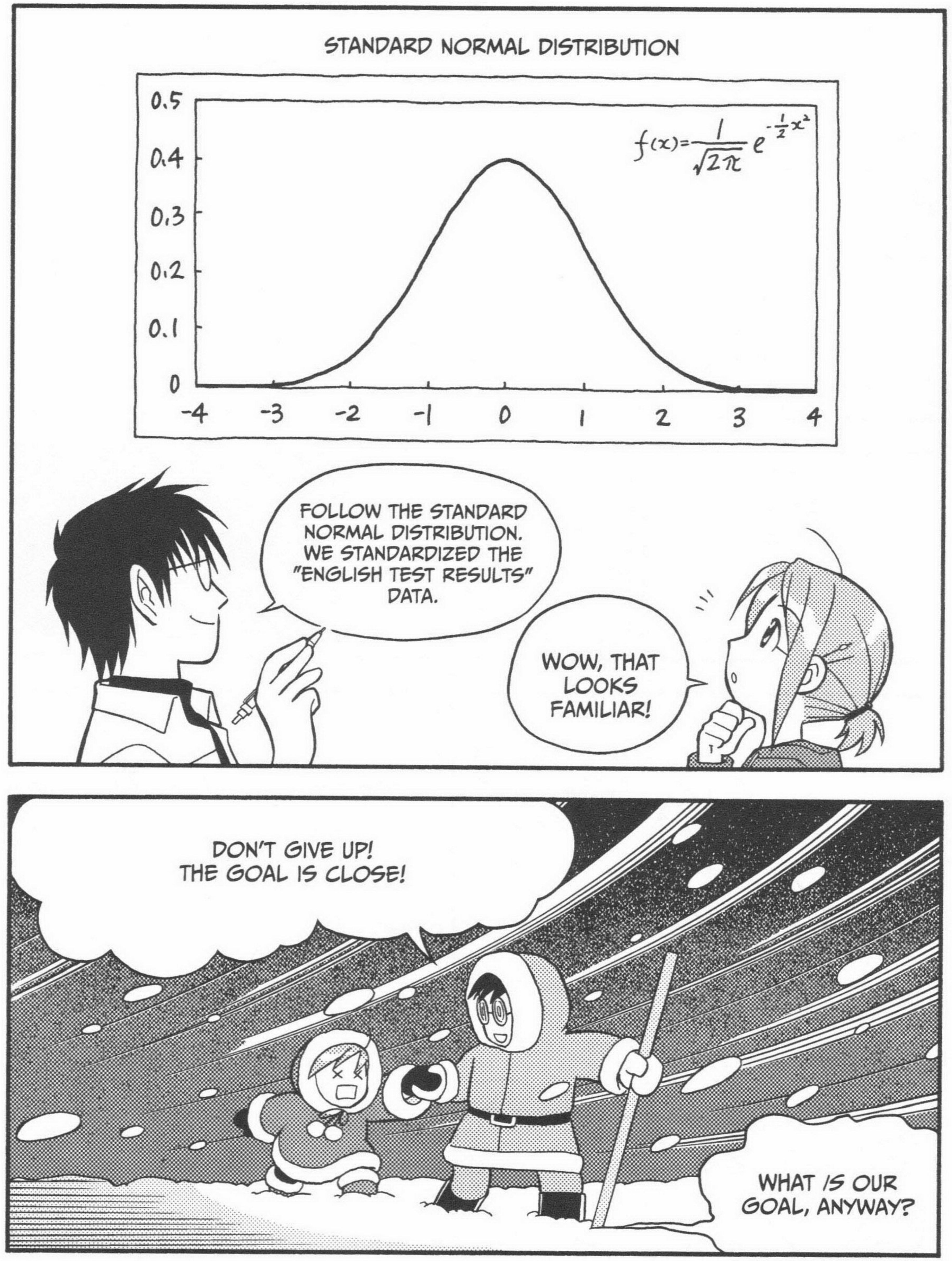
$$F(x) = Erf(x)$$

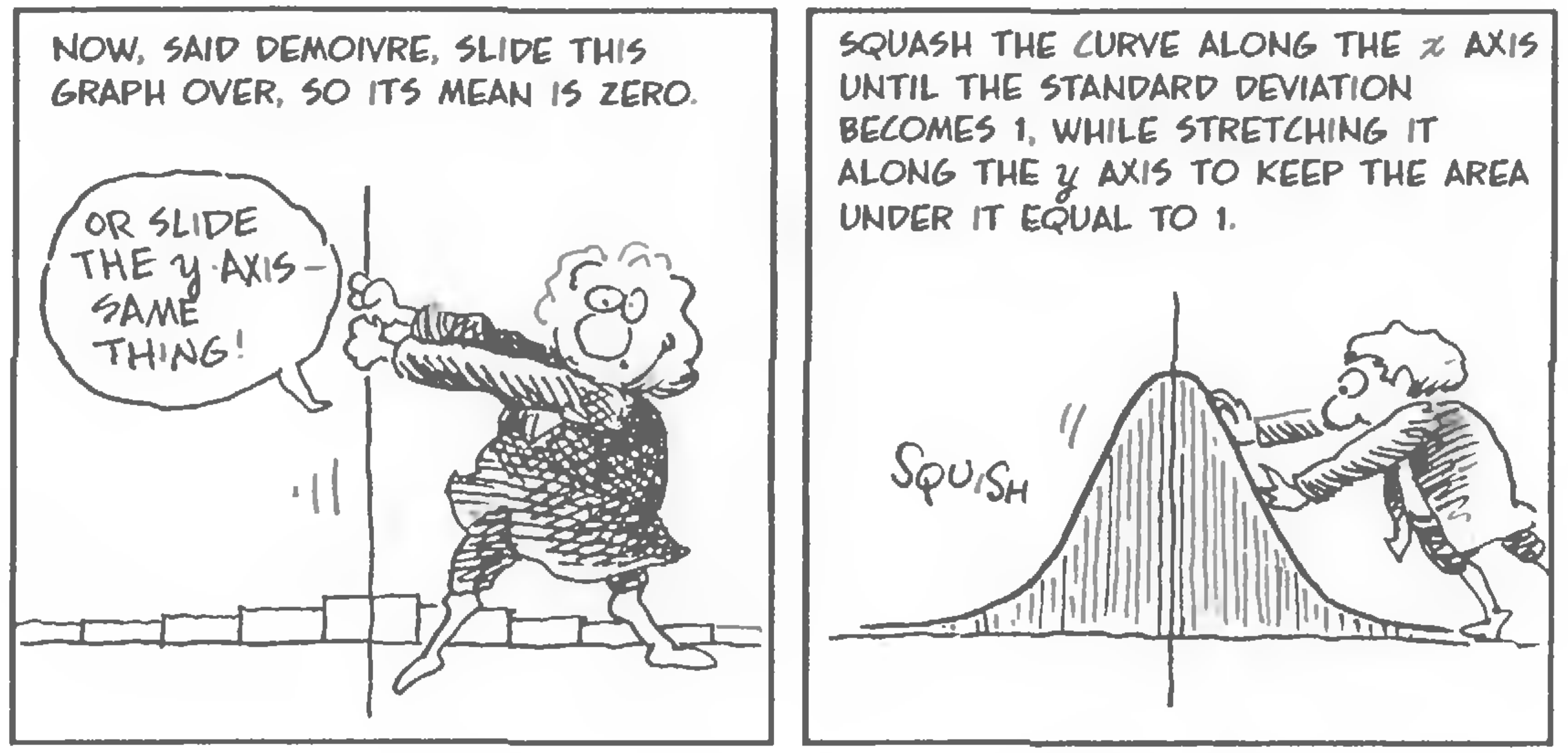


Bell curve



Error function





THE RESULT IS VERY CLOSE TO A ***SMOOTH, SYMMETRICAL, BELL-SHAPED CURVE,*** WHICH DEMOIVRE SHOWED WAS GIVEN BY THE SIMPLE FORMULA:

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

THIS FUNCTION IS CALLED THE **standard normal distribution.**

(e IS A USEFUL MATHEMATICAL CONSTANT APPROXIMATELY EQUAL TO 2.718.)

(CONVINCE YOURSELF THAT THIS FUNCTION REALLY HAS A BELL-SHAPED GRAPH. FOR z FAR FROM ZERO, $f(z)$ IS VERY NEARLY ZERO—IT HAS A BIG DENOMINATOR; IT'S SYMMETRICAL, SINCE $f(z) = f(-z)$, AND IT HAS A MAXIMUM AT $z = 0$.)

THE DISTRIBUTION IS CALLED THE ***STANDARD*** NORMAL BECAUSE ALL THAT SQUASHING AND STRETCHING WAS SPECIALLY ARRANGED TO GIVE IT THESE SIMPLE PROPERTIES, WHICH WE PRESENT WITHOUT PROOF:

$$\mu = 0$$
$$\sigma = 1$$

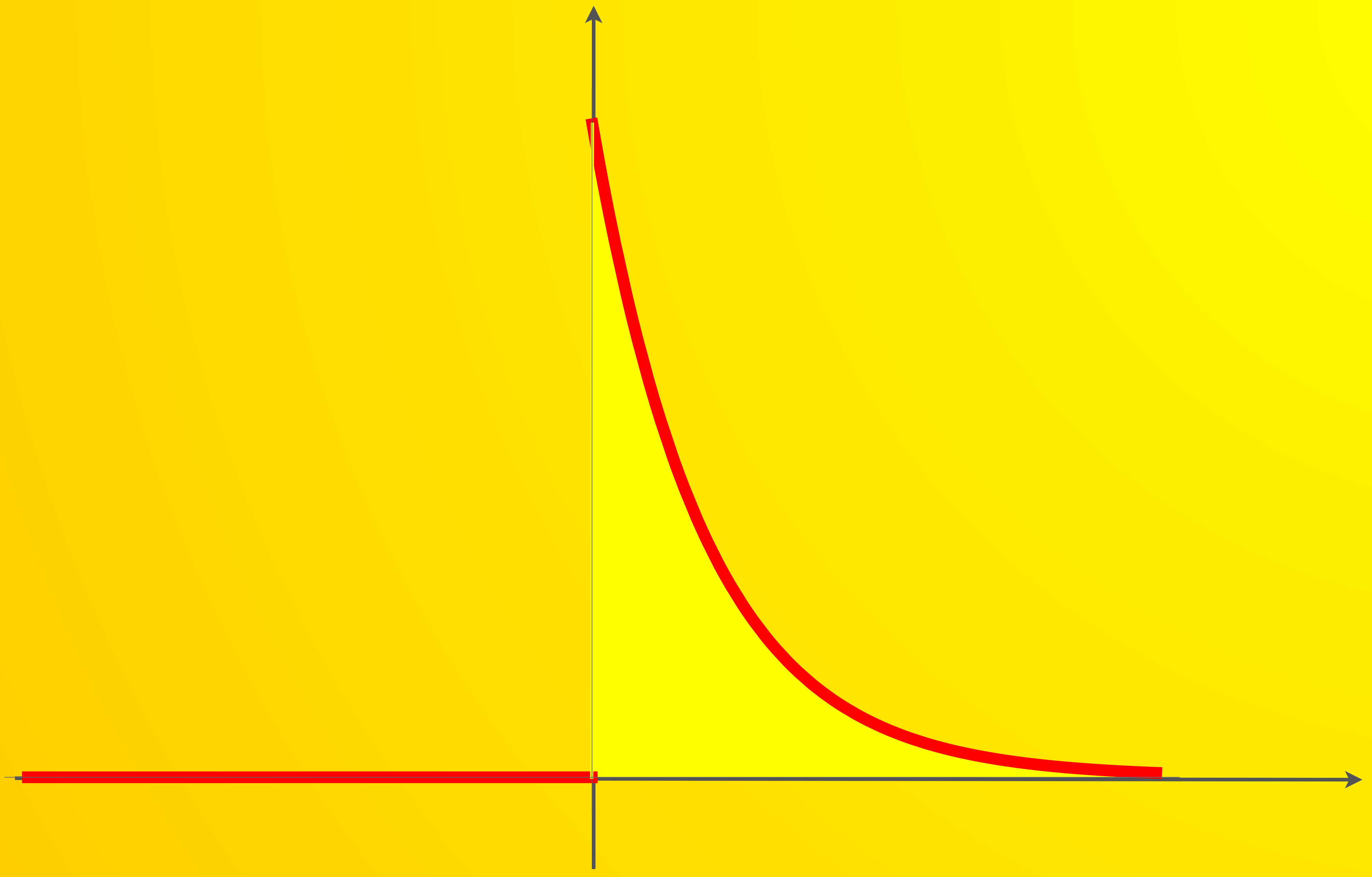
IS IT A PDF?

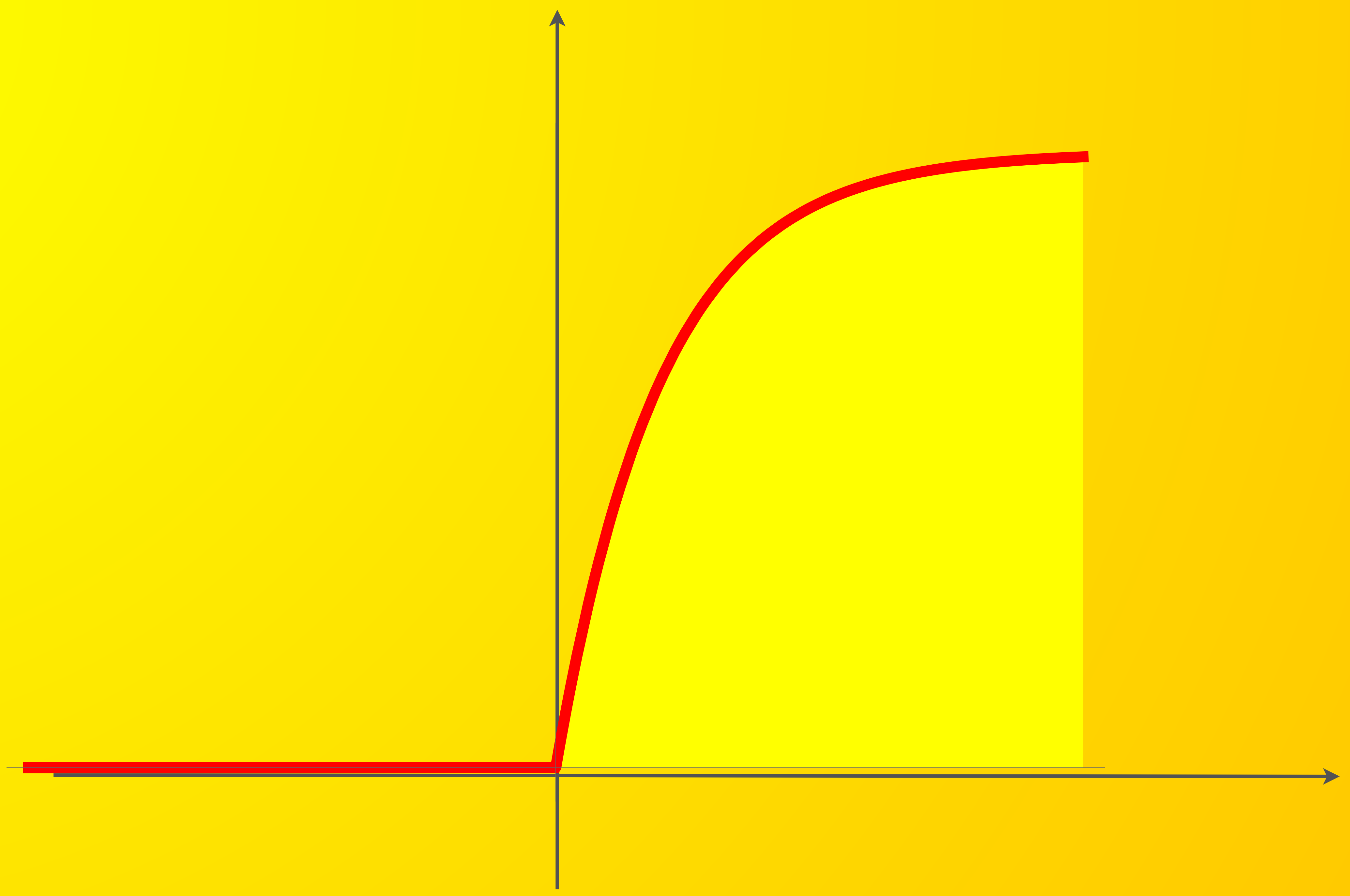
$$\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = 1?$$

IS IT A PDF?

$$f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$F(x) = 1 - e^{-x}$$





Data -> Functions

1000 data points

smoothed PDF

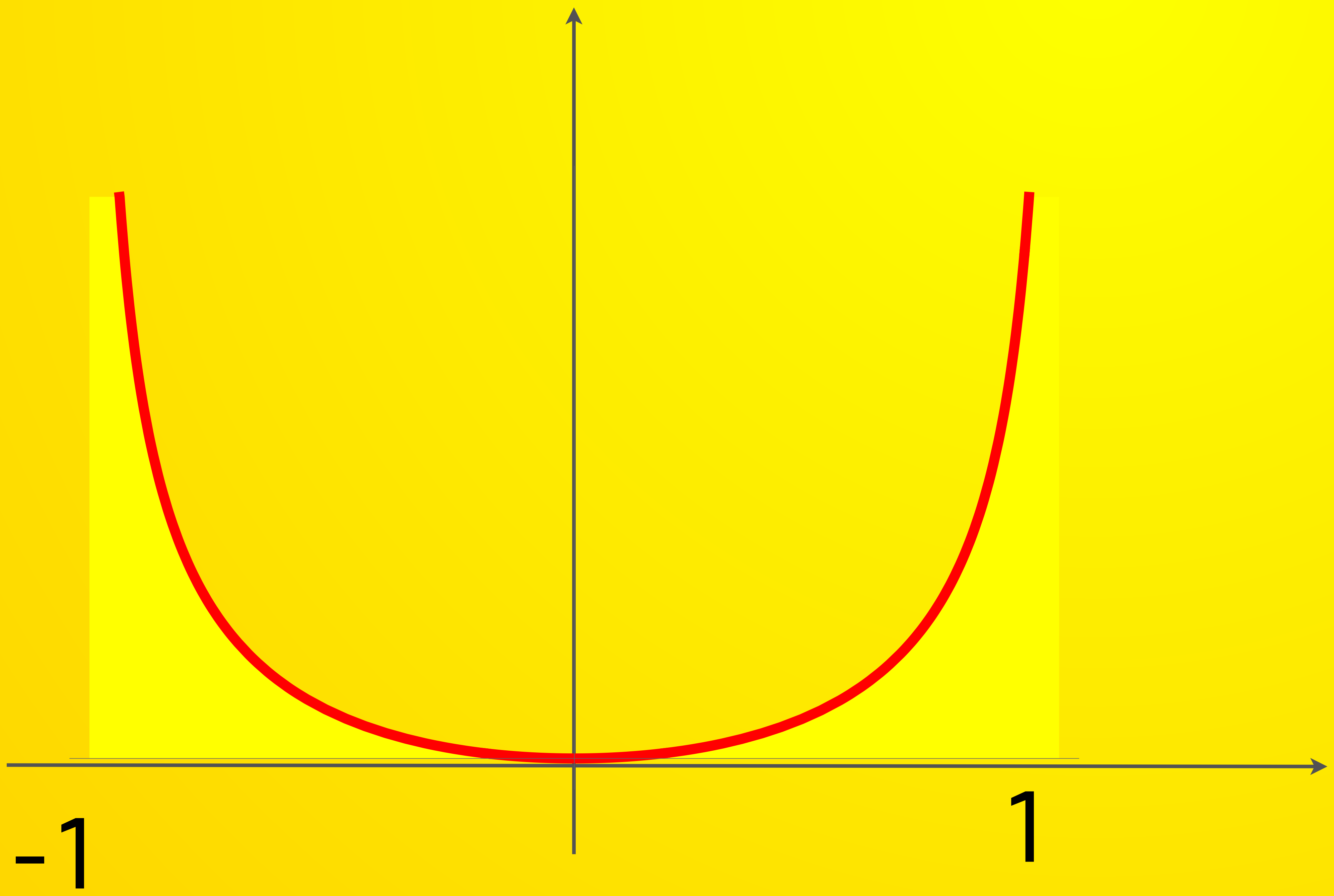
Histogram

ARC SIN

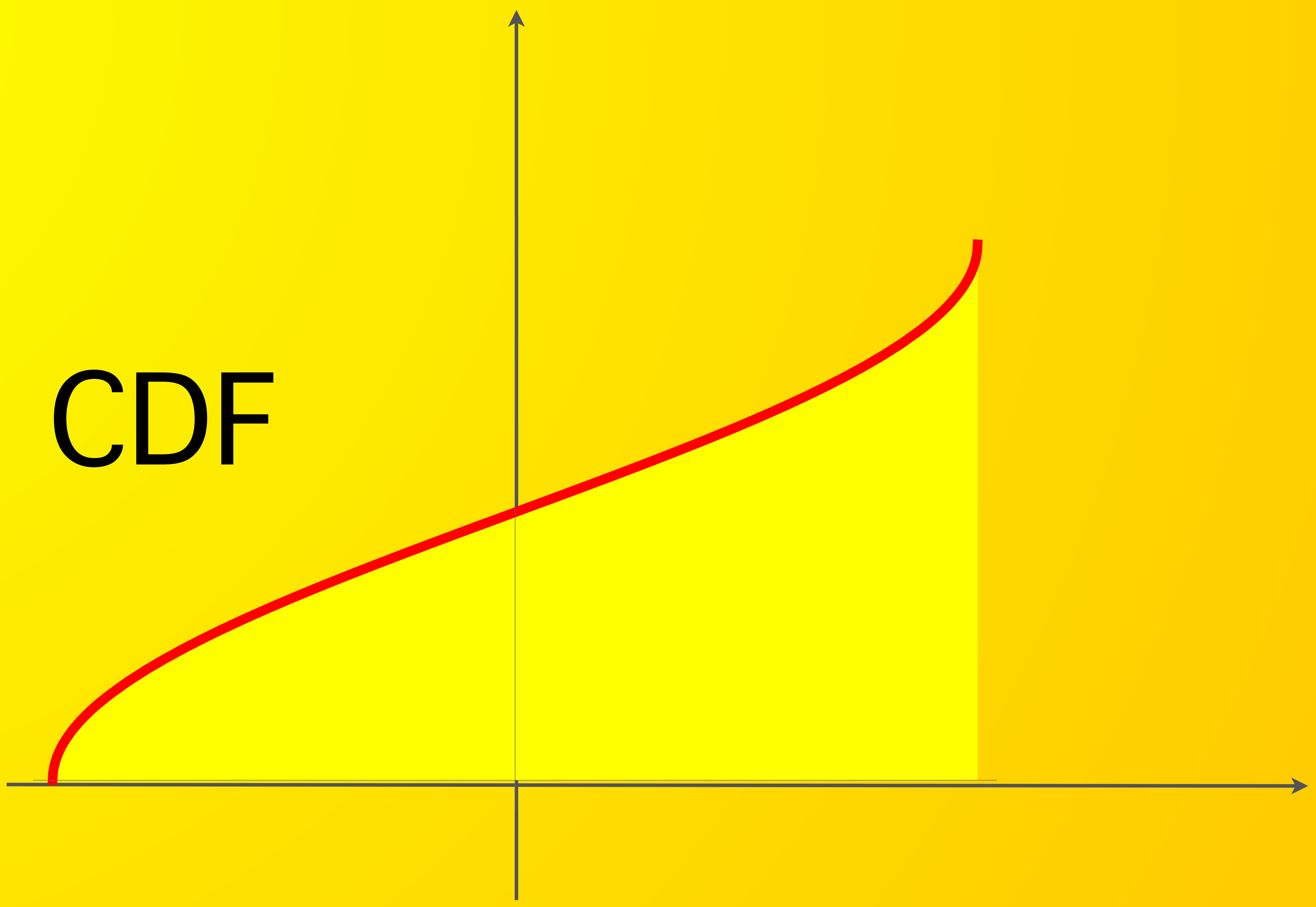
$$f(x) = \frac{1}{\pi\sqrt{1 - x^2}}$$

$$F(x) = \arcsin(x)/\pi$$

PDF



CDF



Data -> Functions

1000 data points

smoothed PDF

Histogram

The Mean

$$\int_{-\infty}^{\infty} xf(x)\,dx = m$$

Moments

$$M_0 = \int_{-\infty}^{\infty} f(x)\, dx = 1$$

$$M_1 = \int_{-\infty}^{\infty} x f(x)\, dx = m$$

Mean

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

n-th moment

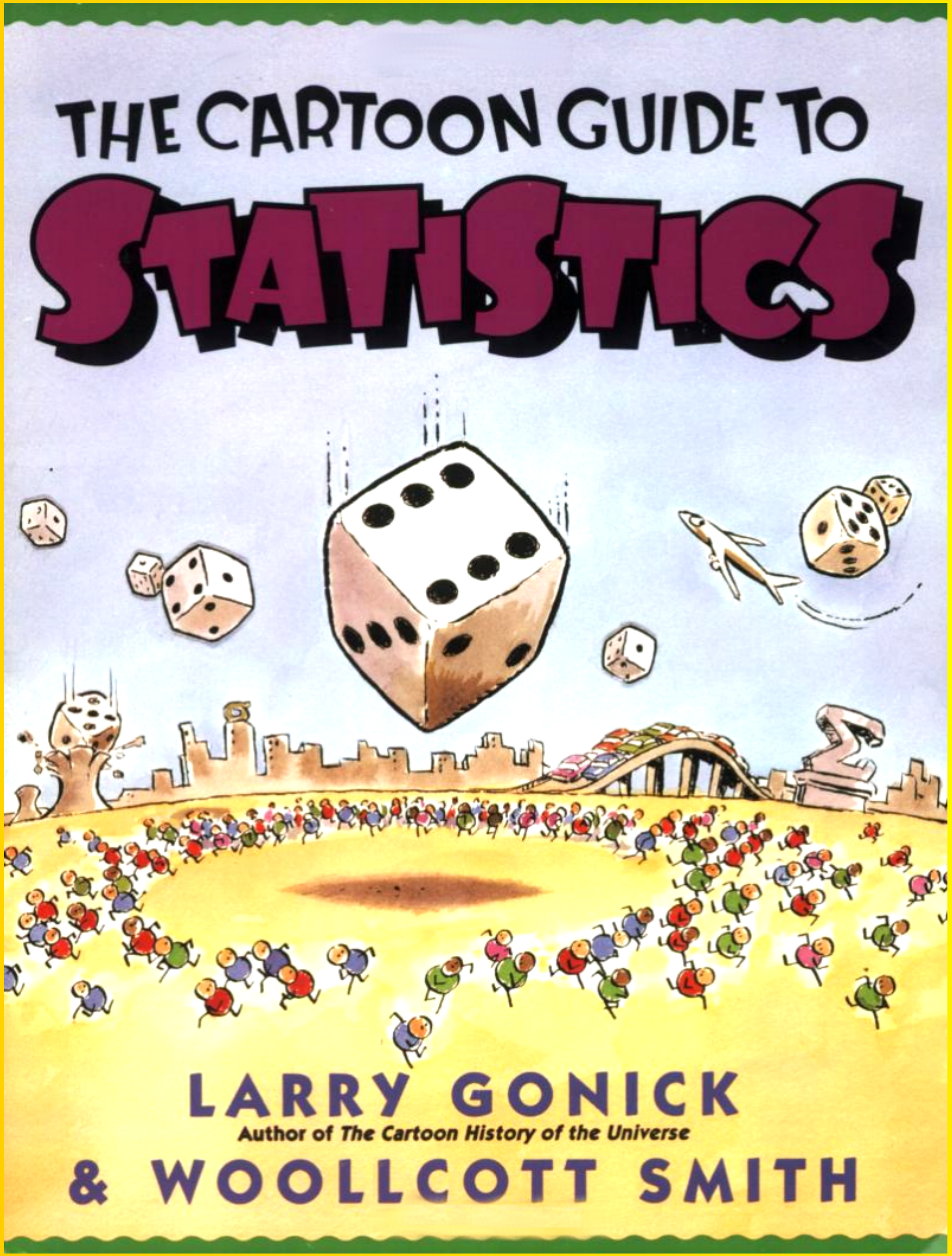
$$\mathrm{Var}(f) = \int_{-\infty}^{\infty} (x - m)^2 f(x)\, dx$$

$$\mathrm{Var}(f) = M_2 - M_1^2$$

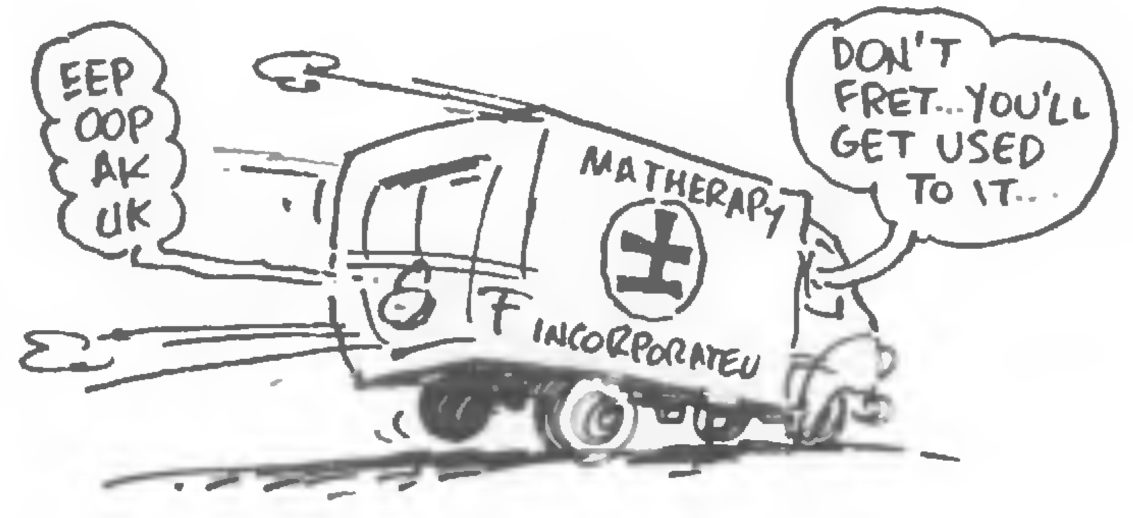
Variance

$$\sigma(f) = \sqrt{\mathrm{Var}(f)}$$

Standard Deviation



ALTHOUGH THE NOTATION MAY BE UNFAMILIAR, ALL IT MEANS IS AN *AREA*.. THE INTEGRAL SIGN ITSELF IS A STRETCHED "S," FOR SUM, WHICH THE INTEGRAL, IN SOME SENSE, IS.



AS A SUMLIKE SOMETHING, THE INTEGRAL SERVES TO DEFINE THE

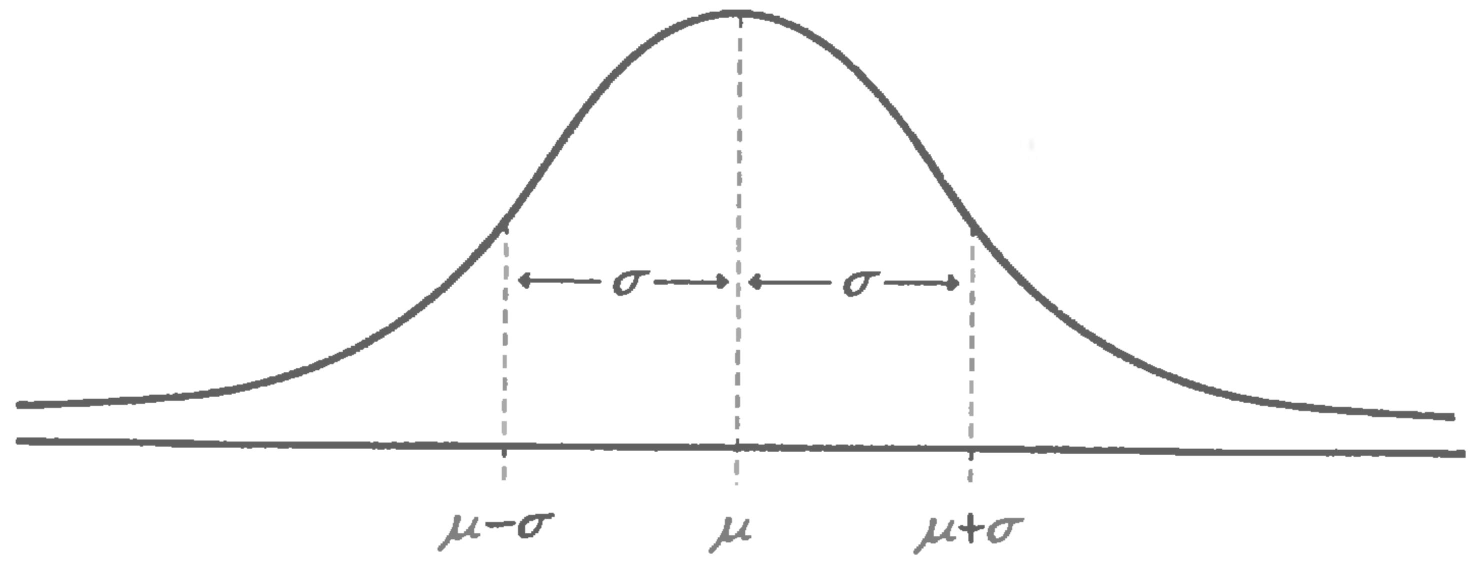
MEAN AND VARIANCE of a continuous random variable.

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx \qquad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

BY ANALOGY WITH THE DISCRETE FORMULAS:

$$\mu = \sum_{\text{all } x} x\,p(x) \qquad \sigma^2 = \sum_{\text{all } x} (x-\mu)^2 p(x)$$

ALTHOUGH IT MAY NOT BE OBVIOUS FROM THE FORMULAS, THESE DEFINITIONS OF MEAN AND VARIANCE ARE ENTIRELY CONSISTENT WITH THEIR ROLE AS CENTER AND AVERAGE SPREAD OF THE PROBABILITIES GIVEN BY THE DENSITY $f(x)$. THE PICTURE TO KEEP IN MIND IS THIS:



Central Moments

Mean

$$M_0 = \int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$M_1 = \int_{-\infty}^{\infty} xf(x)\,dx = m$$

$$\mu_n = \int_{-\infty}^{\infty} (x-m)^n f(x)\,dx$$

n-th central moment

μ_n/σ^n standardized

Skewness $\mu_3(f)/\sigma(f)^3$

$\mathrm{Var}(f) = C_2$ Variance

Kurtosis $\mu_4(f)/\sigma(f)^4$

$\sigma(f) = \sqrt{\mathrm{Var}(f)}$ Standard Deviation

Practice

$$f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

A

$$M_1 = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

B

$$M_2 = \int_{-\infty}^{\infty} x^2\, f(x)\, dx$$

C

$$\mathrm{Var}(f) = M_2 - M_1^2$$

Jam

$$f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Compute

$$\int_{-\infty}^{\infty} x^3 f(x)\, dx$$

For the standard exponential distribution

The End