Probability density functions.

Consider a huge – maybe even uncountably infinite – data set X, whose values all belong to some interval (c,d) of real numbers. (It might be that $c=-\infty$ and/or $d=+\infty$.)

Pick a random sample of size n from X; compute the mean $\bar{x}$ and std. dev. s of the sample. Also draw an RFD histogram for this sample.

Repeat with larger and larger sample sizes n, and narrower and narrower bin widths. Then:

(a) The histograms will converge to some region, bounded above by some function $f(x)$;

(b) The sample means $\bar{x}$ and std. devs. s will converge to numbers μ and σ, respectively.

Definitions

(a) We call f the probability density function, or pdf, for X.

(b) We call μ and σ the (population) mean and standard deviation, respectively, of X.

[See the histograms at the end of these notes.]

FACTS about pdf's (think about these):

If f is a pdf with domain (c,d), then:

(1) $f(x) \geq 0 \quad \forall\, x \in (c,d)$,

(2) For any numbers a and b with $c \leq a \leq b \leq d$, we have

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

the probability that a randomly selected point in X lies in (a,b).

(3) For any single point x_0 in (c,d),

$$P(X = x_0) = \int_{x_0}^{x_0} f(x)\,dx = 0.$$

As a consequence, "pdfs don't care about endpoints," meaning

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

always.

(4) (Formulas for μ and σ.) The grouped data formulas

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \dots + x_k f_k}{n}$$

and

$$s = \sqrt{\frac{f_1(x_1-\bar{x})^2 + f_2(x_2-\bar{x})^2 + \ldots + f_k(x-x_k)^2}{n-1}}$$

become, through the above RFD $\rightarrow$ pdf process,

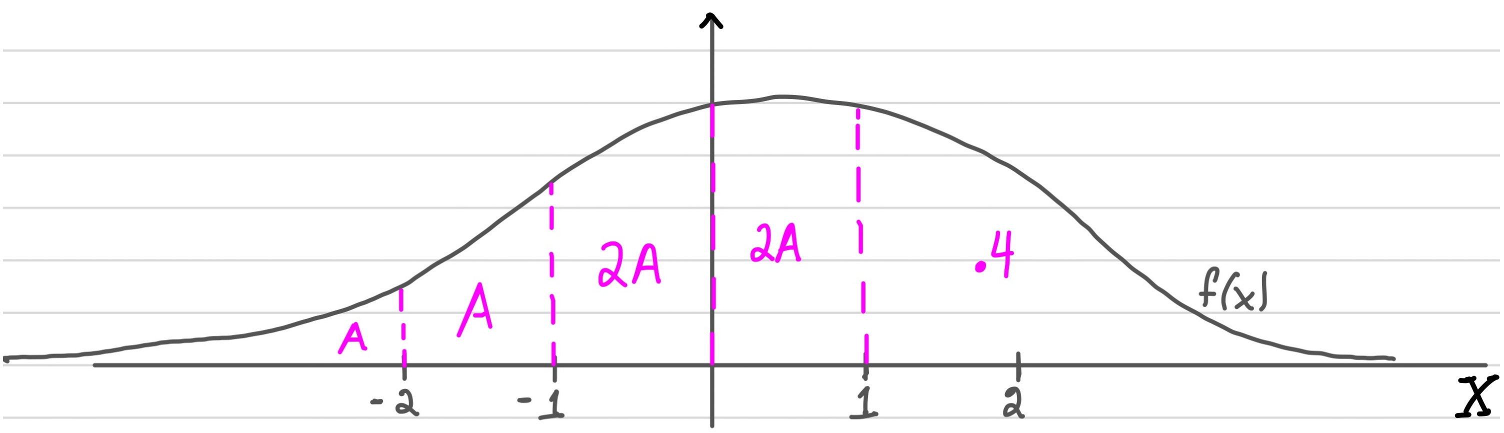
$$\mu = \int_c^d x f(x)\,dx, \quad \text{and}$$

$$\sigma = \sqrt{\int_c^d (x-\mu)^2 f(x)\,dx}\,.$$

(5) $\int_c^d f(x)\,dx = 1.$

Example. Given the pdf below, with domain $(-\infty, \infty)$, find

(a) The value of A;
(b) $P(-2 < X < 0)$;
(c) The number k such that $P(X > k) = 0.8$.



Solution.

We have

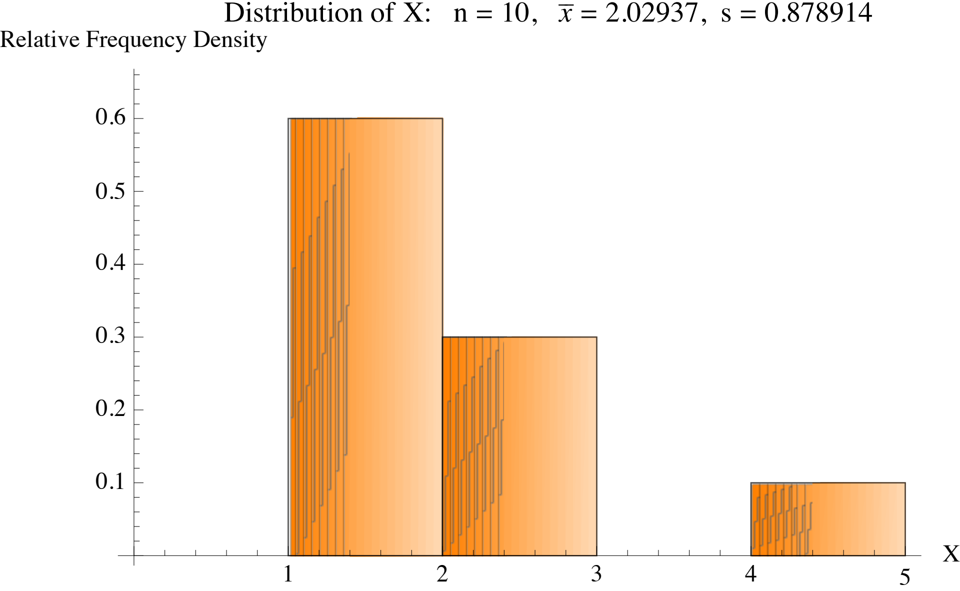
$$A + A + 2A + 2A + .4 = 1$$
$$6A + .4 = 1$$

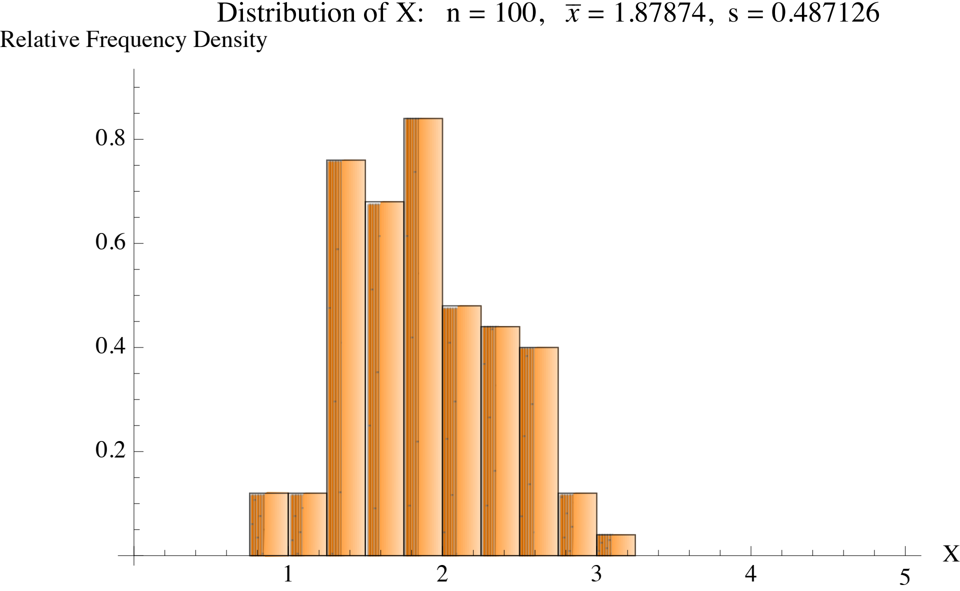
$$6A = 1 - .4 = .6$$
$$A = .1.$$

(b) $P(-2 < X < 0) = A + 2A = 3A = .3.$

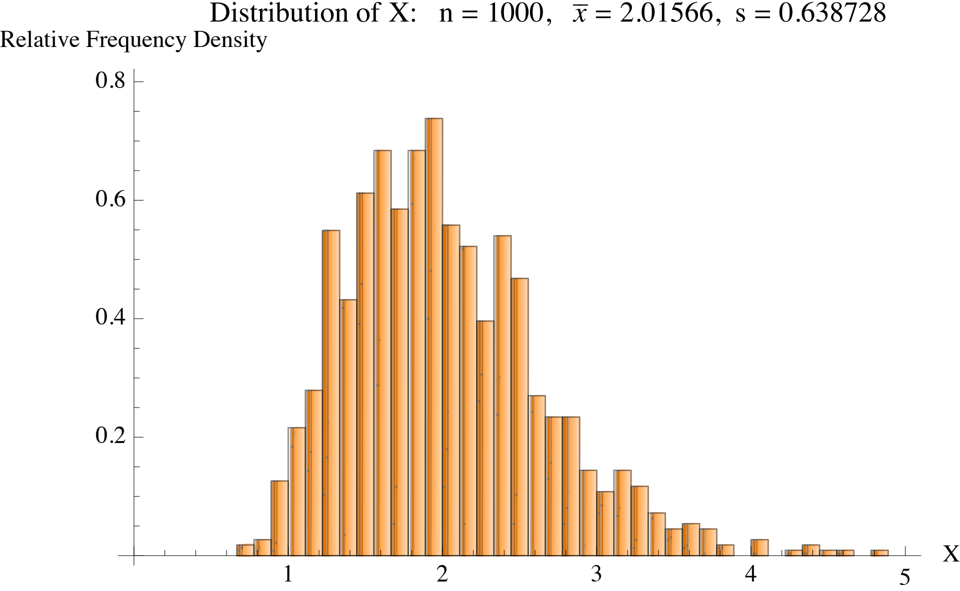
(c) $k = -1$ works, since

$$P(X > -1) = 2A + 2A + .4 = 4A + .4 = .4 + .4$$
$$= .8.$$

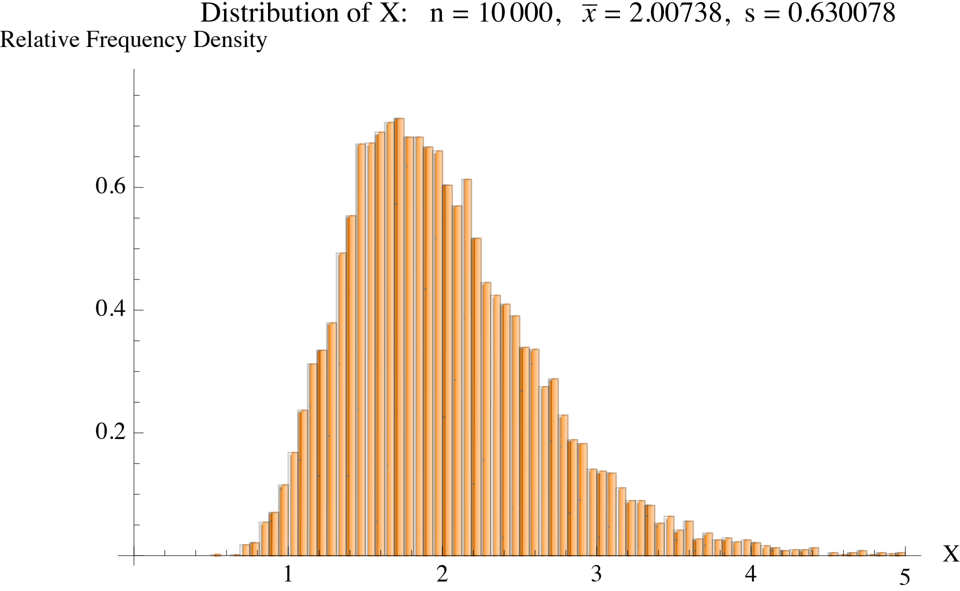

Distribution of X: n = 10, $\bar{x}$ = 2.02937, s = 0.878914
Relative Frequency Density
0.6
0.5
0.4
0.3
0.2
0.1
1
2
3
4
5
X



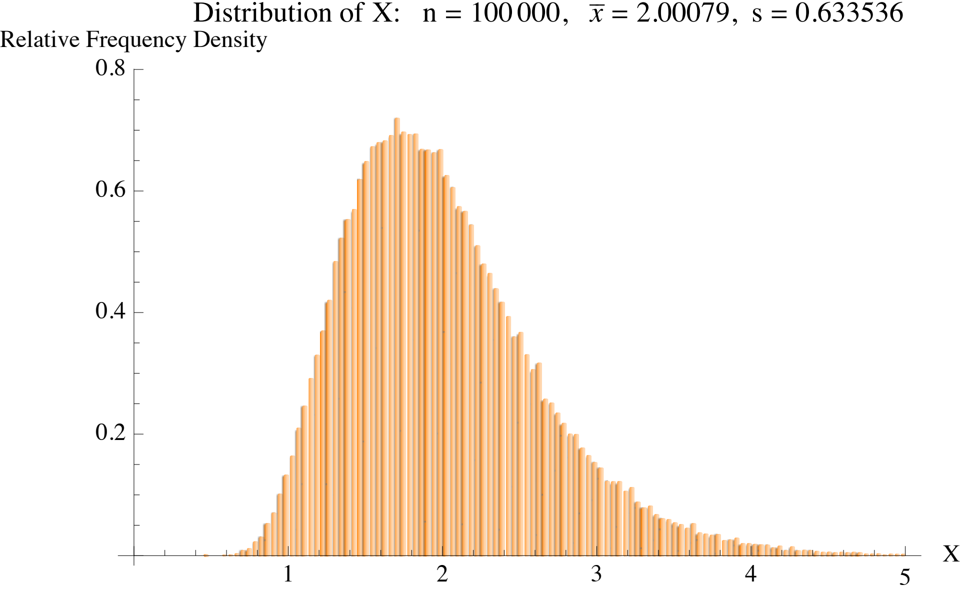
Distribution of X: n = 100, $\bar{x}$ = 1.87874, s = 0.487126
Relative Frequency Density
0.8
0.6
0.4
0.2
1
2
3
4
5
X



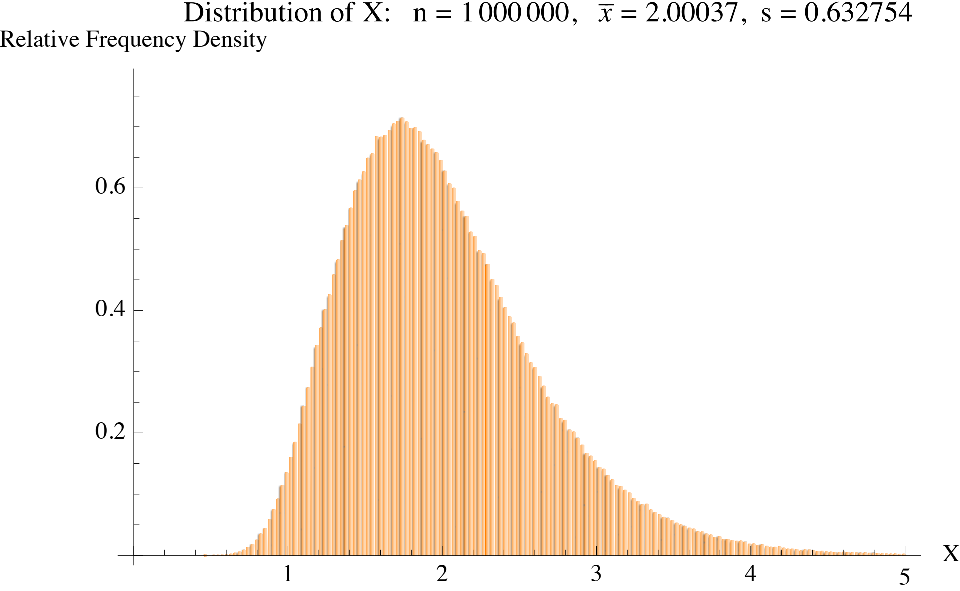
Distribution of X: n = 1000, $\bar{x}$ = 2.01566, s = 0.638728
Relative Frequency Density
0.8
0.6
0.4
0.2
1
2
3
4
5
X



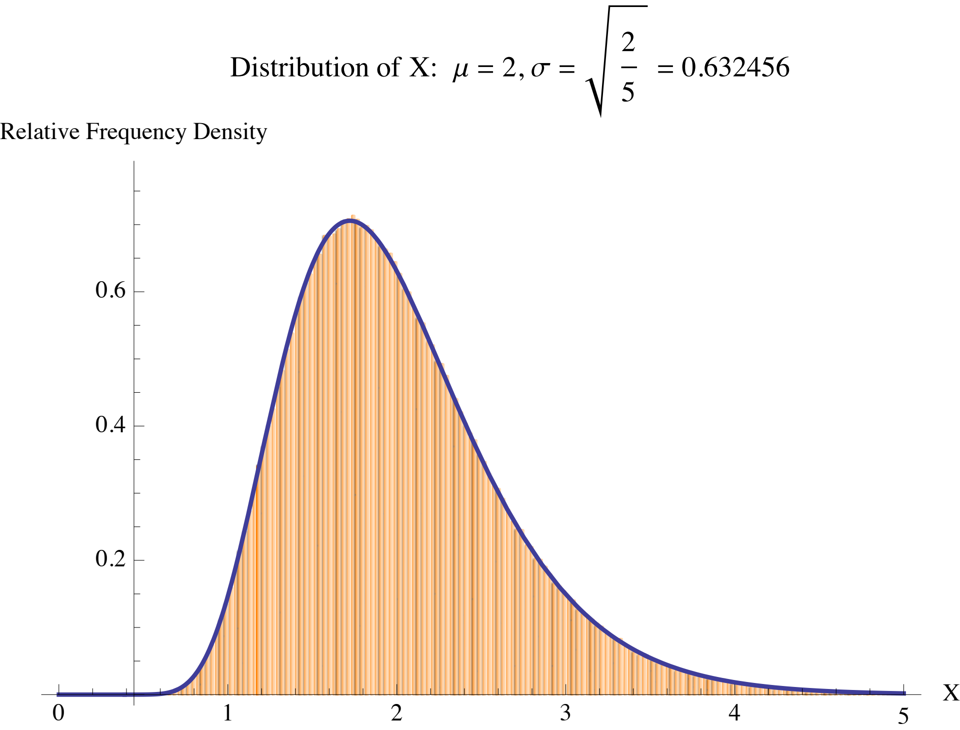
Distribution of X: n = 10 000, $\bar{x}$ = 2.00738, s = 0.630078
Relative Frequency Density
0.6
0.4
0.2
1
2
3
4
5
X



Distribution of X: n = 100 000, x̄ = 2.00079, s = 0.633536
Relative Frequency Density
0.8
0.6
0.4
0.2
1
2
3
4
5
X



Distribution of X: n = 1 000 000, $\bar{x}$ = 2.00037, s = 0.632754
Relative Frequency Density
0.6
0.4
0.2
1
2
3
4
5
X


Distribution of X: $\mu = 2, \sigma = \sqrt{\frac{2}{5}} = 0.632456$
Relative Frequency Density
0.6
0.4
0.2
0
1
2
3
4
5
X