CSCI 3022

intro to data science with probability & statistics

Lecture 12
Feb 23, 2018

1. The normal distribution

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

~13 mult choice = 40
3 free response = 60
100

- **Homework 3** due next Friday, March 2. Suggested **milestones**:
  - Probs 1, and 2, done before the end of the week.
  - Probs 3 and 4 done next week.
  - Midterm week format: 5 problems but *you choose only 4*.
- **Midterm** next week. Weds, Feb 28, 6:30-7:50 PM. **HUMN 1B50**.
  - Start of the course up through variance.
- Midterm review *in class* next Monday, Feb 26.
  - Bring yer questions!

Last time: E[Uniform]

- **Suppose** $X \sim U[\alpha, \beta]$. Find $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$= \int_{\alpha}^{\beta} x\, f(x)\, dx$$

$$= \int_{\alpha}^{\beta} x \frac{1}{\beta - \alpha}\, dx$$

$$\frac{x^2}{2}\Big|_{\alpha}^{\beta} \frac{1}{\beta-\alpha}$$

$$= \frac{\beta^2 - \alpha^2}{2(\beta-\alpha)}$$

$$= \frac{\cancel{(\beta-\alpha)}(\beta+\alpha)}{2\cancel{(\beta-\alpha)}}$$

$$= \boxed{\frac{\beta+\alpha}{2}}$$

Var(Uniform) ?

- **Suppose** $X \sim U[\alpha, \beta]$. Find $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.

$$\frac{\beta^2 + 2\alpha\beta + \alpha^2}{4}$$

$$\mathrm{Var}(X) = \underline{E[X^2]} - \underline{E[X]^2} \qquad E[X] = \frac{\beta+\alpha}{2} \Rightarrow E[X]^2 = \left(\frac{\beta+\alpha}{2}\right)^2$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)$$

$$= \int_{\alpha}^{\beta} x^2 \frac{1}{\beta-\alpha}\, dx$$

$$= \frac{1}{\beta-\alpha} \frac{1}{3} x^3 \Big|_{\alpha}^{\beta}$$

$$\frac{1}{3} \frac{1}{\beta-\alpha} \left(\beta^3 - \alpha^3\right)$$

$$= \frac{\cancel{(\beta-\alpha)}\left(\beta^2 + \cancel{2}\alpha\beta + \alpha^2\right)}{3\ \cancel{(\beta-\alpha)}}$$

$$= \frac{\beta^2 + \cancel{2}\alpha\beta + \alpha^2}{3}$$

$$\mathrm{Var}(X) = \frac{\beta^2 + \cancel{2}\alpha\beta + \alpha^2}{3} - \frac{\beta^2 + 2\alpha\beta + \alpha^2}{4} = \frac{1}{12}(\beta-\alpha)^2 \checkmark$$

std dev $\frac{1}{\sqrt{12}} |\beta - \alpha|$ ✓

What happened with this derivation in class?

We wanted $E[X^2]$, which is $\int_\alpha^\beta x^2 \frac{1}{\beta-\alpha}\,dx$

Computing this integral, $E[X^2] = \frac{1}{\beta-\alpha} \frac{x^3}{3}\Big|_\alpha^\beta$

$$= \frac{1}{3}\,\frac{\beta^3-\alpha^3}{\beta-\alpha}$$

In class, a mistake!

I factored $\beta^3-\alpha^3 = (\beta-\alpha)(\beta^2+2\alpha\beta+\alpha^2)$

In fact, $\beta^3-\alpha^3 = (\beta-\alpha)(\beta^2+\alpha\beta+\alpha^2)$

same

Yikes! correct factorization has a 1, not a 2.

$$= \frac{1}{3}\,\frac{\cancel{(\beta-\alpha)}(\beta^2+\alpha\beta+\alpha^2)}{\cancel{(\beta-\alpha)}}$$

$$= \frac{1}{3}(\beta^2+\alpha\beta+\alpha^2)$$

Thus, $E[X^2] - E[X]^2 = \frac{1}{3}(\beta^2+\alpha\beta+\alpha^2) - \frac{1}{4}(\beta^2+2\alpha\beta+\alpha^2)$

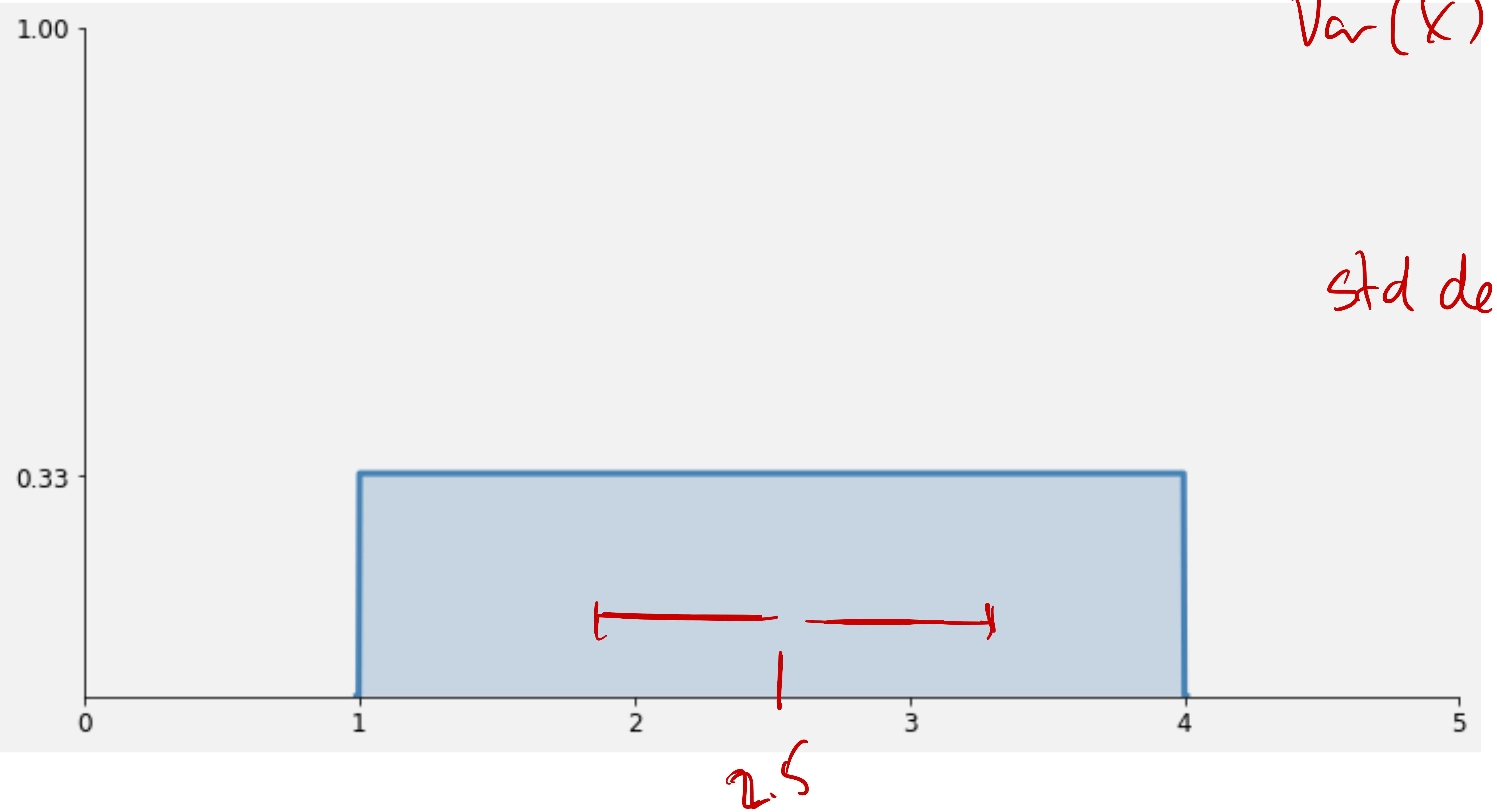
$$= \frac{1}{12}(\beta-\alpha)^2 \checkmark$$

Mean & Variance: Uniform

- **Theorem**: Let $X \sim U[\alpha, \beta]$. Then: $\mathrm{E}[X] = \dfrac{\alpha + \beta}{2}$ $\quad \mathrm{Var}(X) = \dfrac{(\beta - \alpha)^2}{12}$

$\mathrm{Var}(X) = \dfrac{3^2}{12} = \dfrac{9}{12} = \dfrac{3}{4}$

std dev $= \dfrac{\sqrt{3}}{2}$



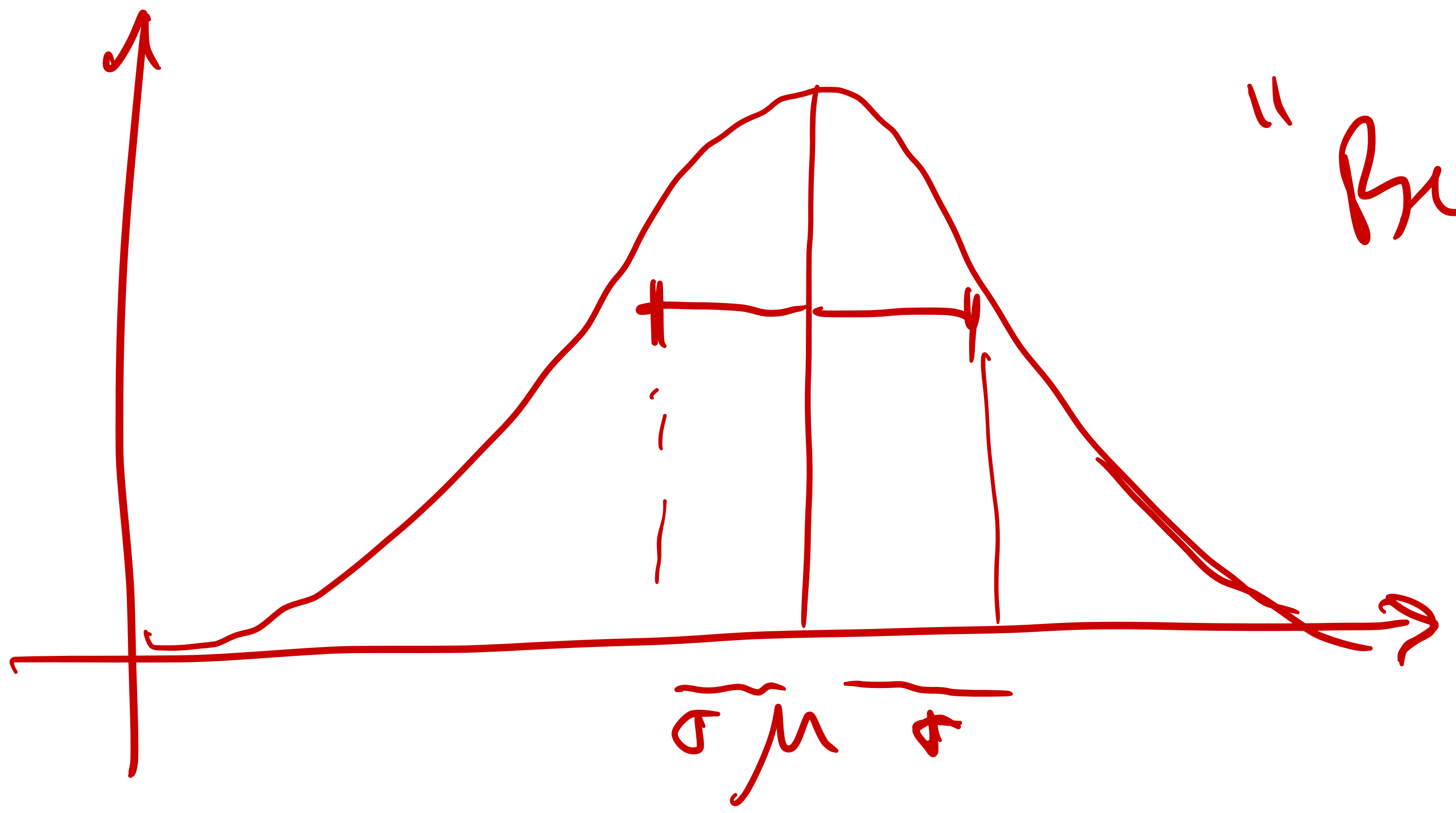
The normal distribution

Gauss

- The **normal distribution** (aka the Gaussian distribution) is probably the most important distribution in all of probability and statistics!
- Many populations have distributions that are well-approximated by an appropriate normal distribution.
- **Examples:** height, weight, and other physical characteristics, scores on various tests, etc.

Correct μ

correct σ^2



"Bell" shape

The normal distribution

- **Definition**: A continuous random variable X is said to have a **normal distribution** with parameters μ and $\sigma > 0$ (or μ and σ^2) if the pdf of X is:

$$\underset{\substack{\uparrow \\ \text{variable}}}{f(x;}\ \underset{\substack{\uparrow\ \uparrow \\ \text{parameters}}}{\mu, \sigma^2)} = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{\mu - x}{\sigma}\right)^2\right]}$$

$$\left(\frac{x-\mu}{\sigma}\right)^2$$

variance

- If a random variable is normally distributed, we say: $X \sim N(\mu, \sigma^2)$

https://academo.org/demos/gaussian-distribution/

The Standard Normal Distribution

- **Definition**: A normal distribution with parameter values $\mu = 0, \ \sigma^2 = 1$ is called the **standard normal distribution**.

- A random variable with this distribution is called a standard normal random variable, and is often denoted by Z. Its PDF is:

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

- We use a special notation to denote the CDF of the standard normal curve:

$$F(z) = \int_{-\infty}^{z} f(s)\,ds = \boxed{\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} = \Phi(z)}$$

The Standard Normal Distribution

- The standard normal distribution *rarely* occurs naturally.
- Instead, it's a **reference distribution** that allows us to learn about *other* (non-standard) normal distributions using a simple formula.
- Recall that for computing probability integrals, having the CDF is just as good as having the PDF. (Can you recall why?)   In the past, when we used paper books, the values of the standard normal CDF could be found in "normal tables" in the back of any probability textbook.

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

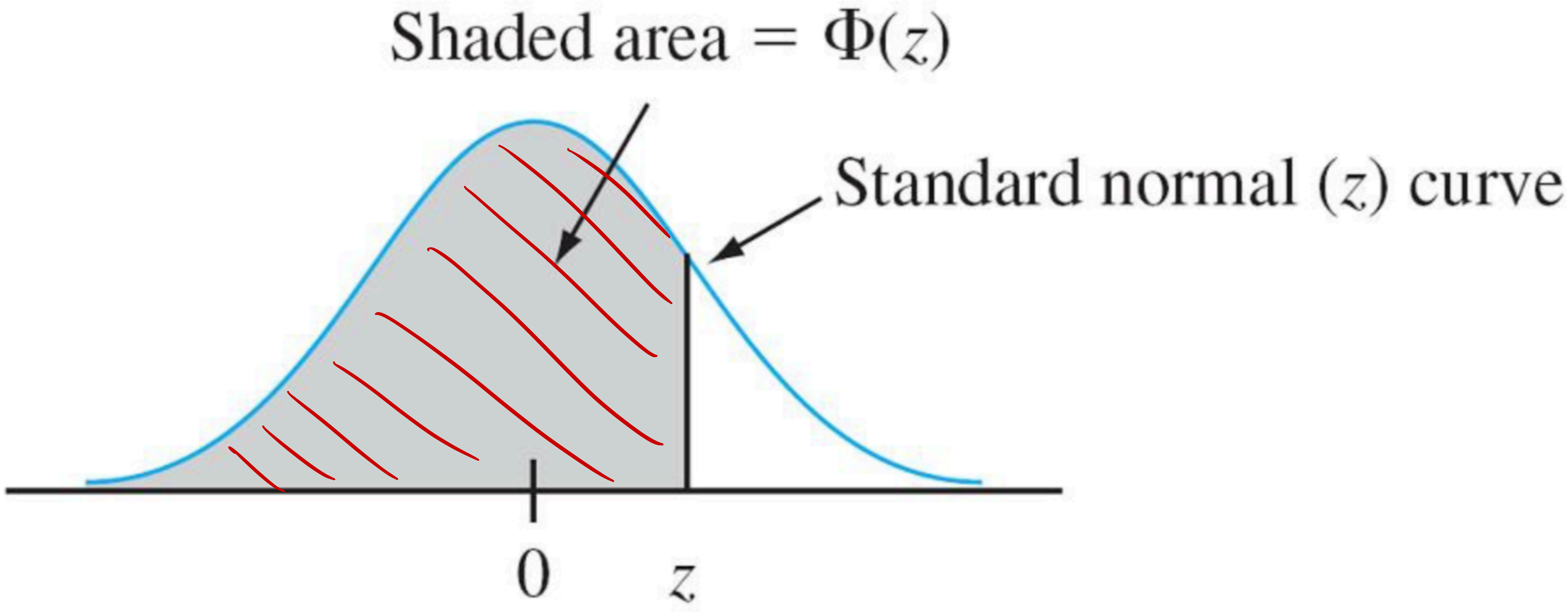
PDF

CDF ϕ

values of a	$\phi(a)$

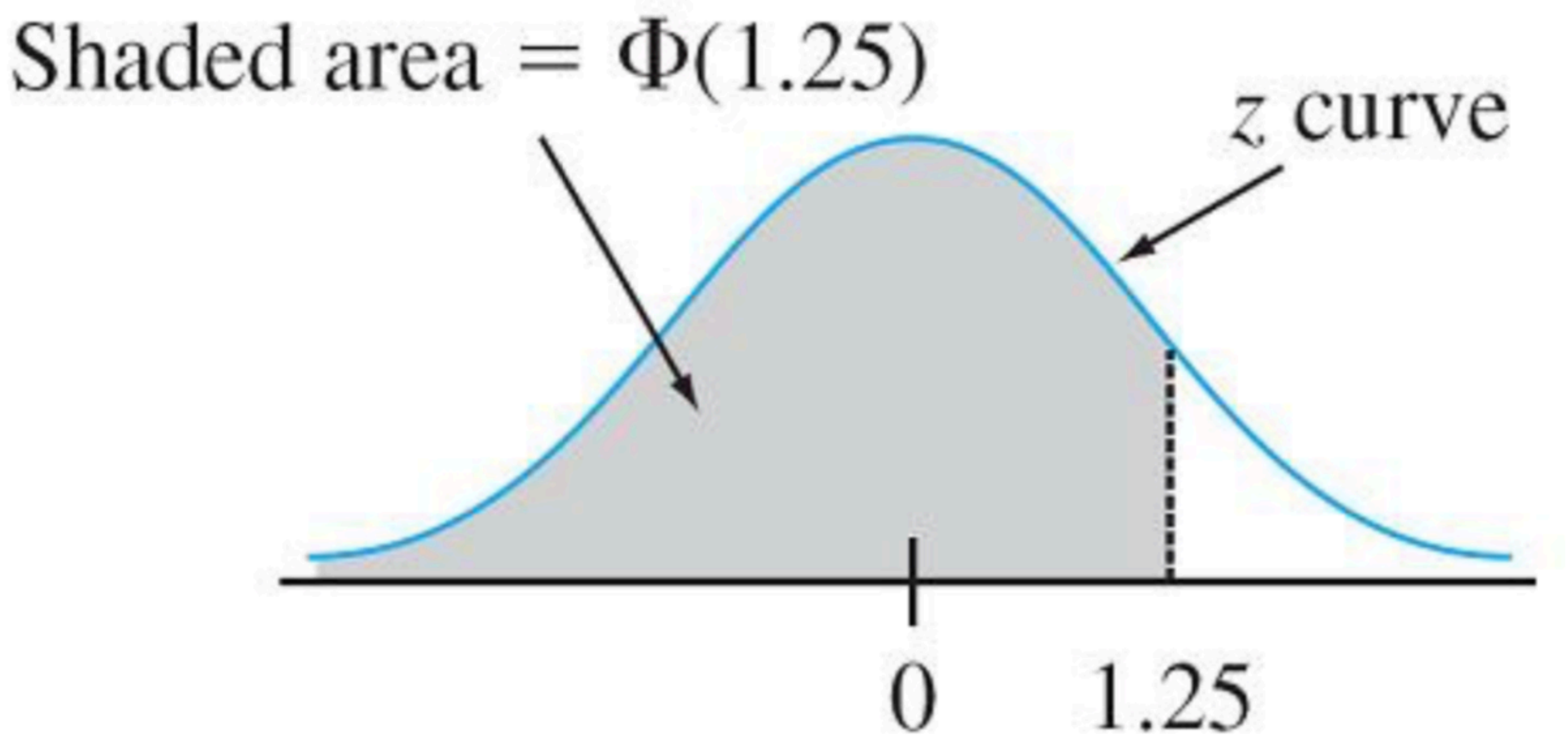
The Standard Normal Distribution

- What's in a standard normal table?



The Standard Normal Distribution

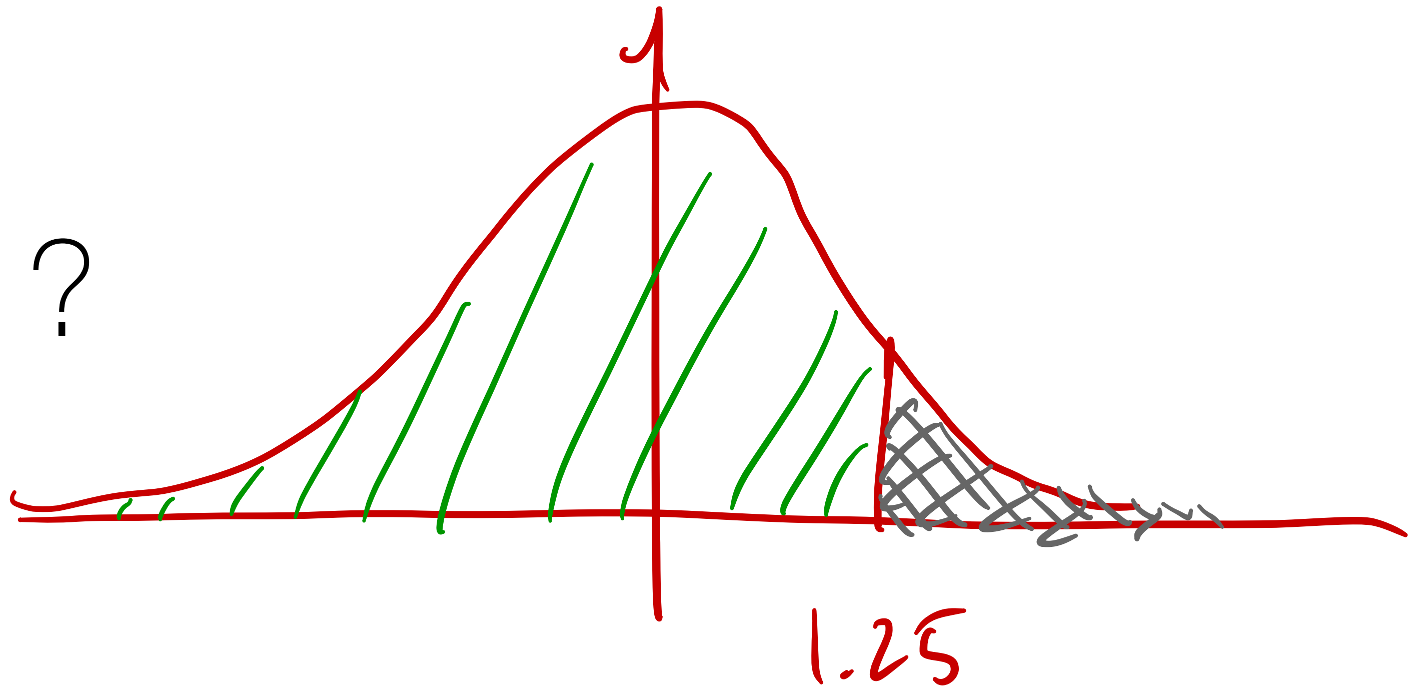
- Example: What is **P(Z≤1.25)** ?



The Standard Normal Distribution

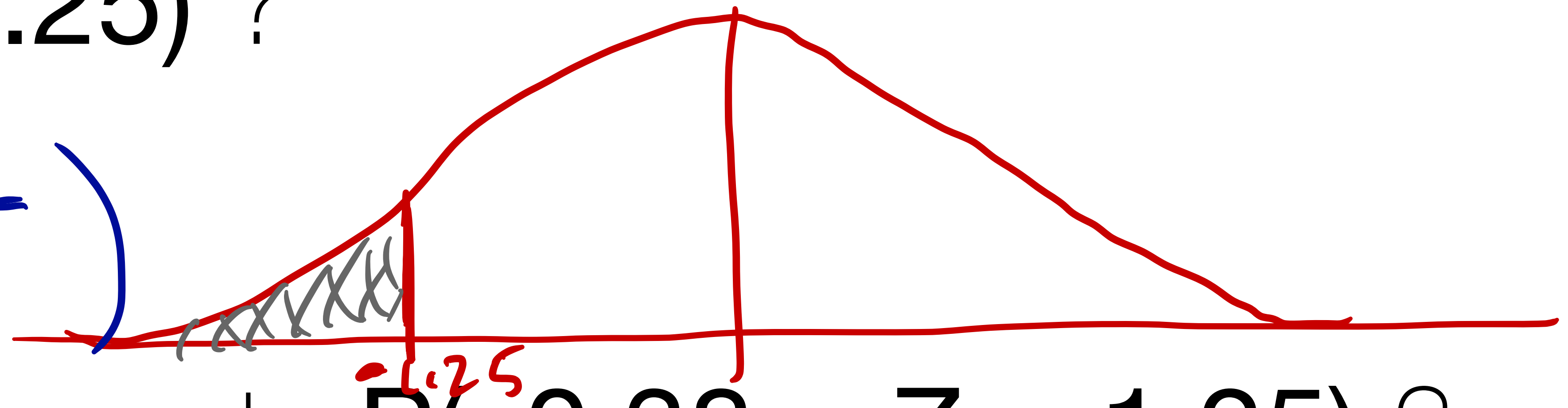
- **Example 1.** What is **P(Z≥1.25)** ?

$1 - \Phi(1.25)$



- **Example 2.** What is **P(Z≤-1.25)** ?

$\Phi(-1.25)$ or $1 - \Phi(1.25)$



- **Example 3.** How can we compute **P(–0.38 ≤ Z ≤ 1.25)** ?

$\Phi(1.25) - \Phi(-0.38)$

Flip it & Reverse it

- **Recall**: what is the 99th percentile of the standard normal distribution?

$$\int_{-\infty}^{P_{99}} f(x)\, dx = 0.99 \qquad \Phi(P_{99}) = 0.99$$

- **Hmmm**…Tables give you a lookup from z to the area under the curve to the left of z. But we have the area and we want z.
- This is the "inverse" problem to P(Z≤z) = 0.99

z	$\Phi(z)$

- How would you use a table?

$\boxed{} \leftarrow 0.99$

- How could you sort this out in Python?

Flip it & Reverse it

- Of course everything that we could look up in a textbook table is built into Python:

```
In [37]:
from scipy import stats
stats.norm.ppf(0.99)
```

```
Out[37]: 2.3263478740408408
```

2.3263 . . . .

- stats.norm has lots of good functions related to normal distributions: pdf, cdf, ppf, etc

… and apply it

- **Notation**: z_α is the value of z under the standard normal distribution that gives a certain "tail" area. In particular, it is the z value such that exactly α area lies to the right of z_α.

- **Hmmm…** so what is the relationship between z_α and the CDF?

$$\Phi(z_\alpha) = 1 - \alpha$$

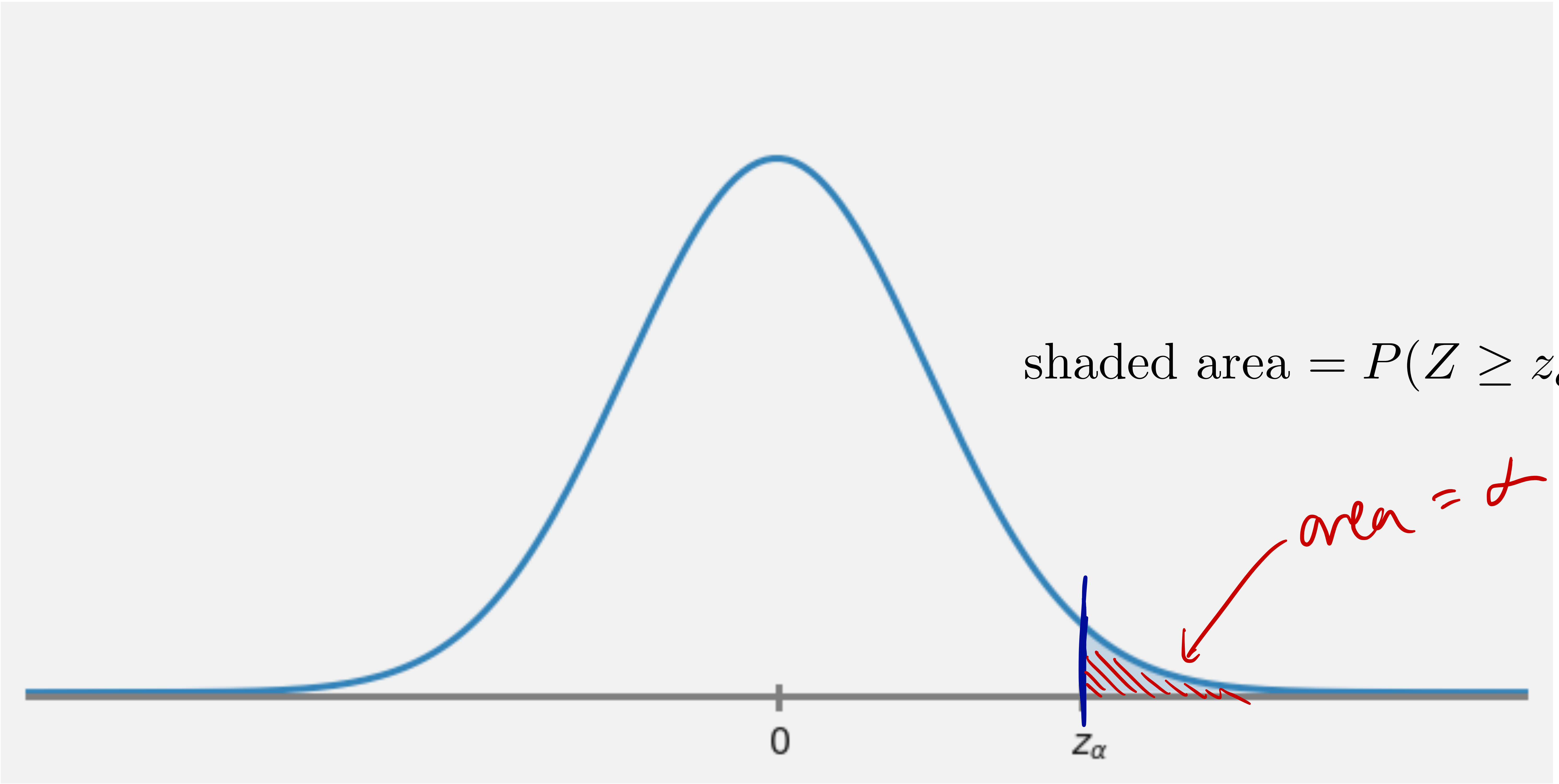
- **Hmmm…** so what is the relationship between z_α and percentiles?

99^{th} percentile $= z_{0.01}$

50^{th} percentile $= z_{0.5}$
(median)

Critical values

- **Notation**: z_α is the value of z under the standard normal distribution that gives a certain "tail" area. In particular, it is the z value such that exactly α area lies to the right of z_α.



Nonstandard normals

- Normal distributions that are not standard can be turned into standard normals so, so, so easily!
- **Proposition**: if X is a normal distribution with mean mu and standard deviation sigma, then Z is a standard normal distribution if:

μ

σ

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

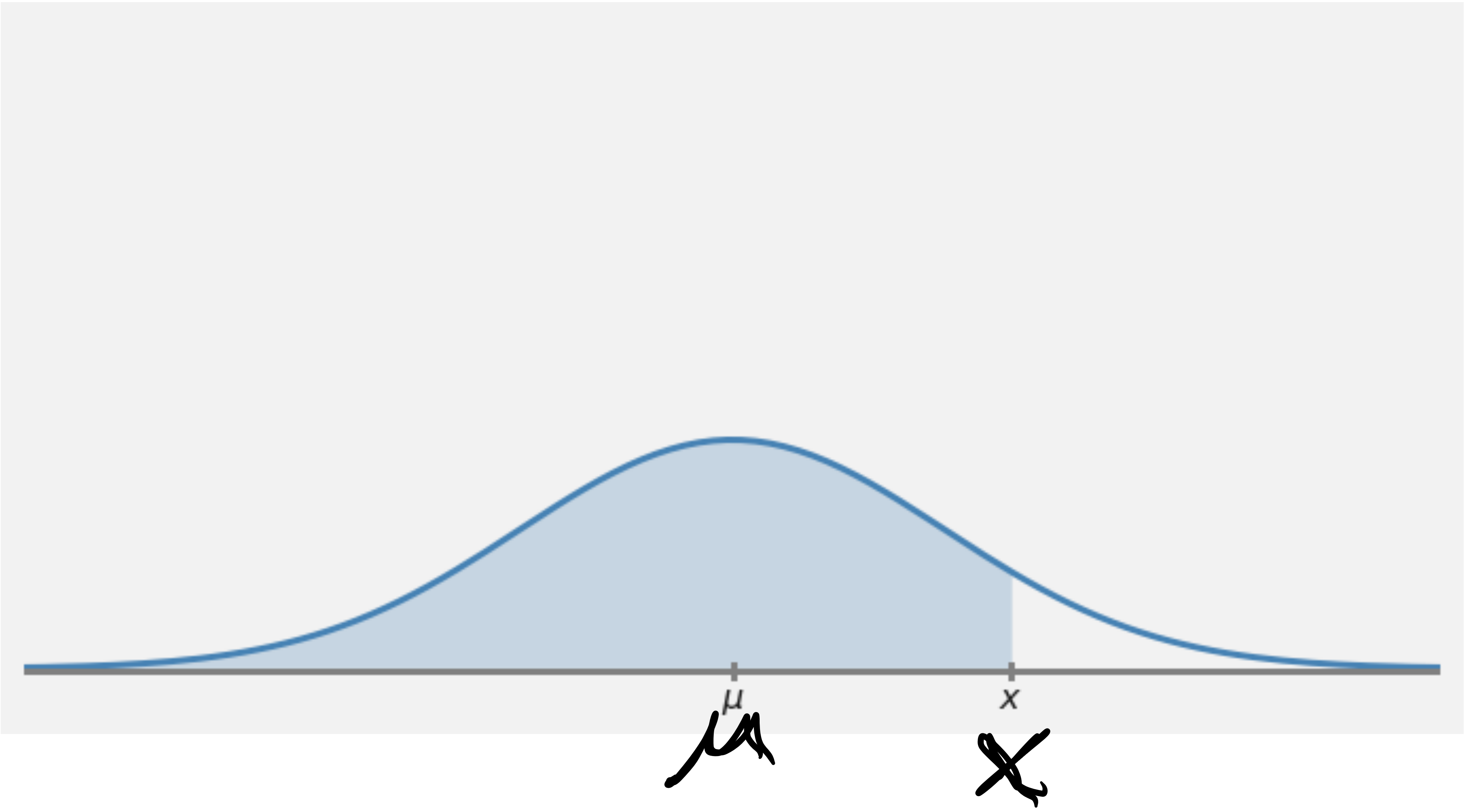
Box Muller Transform

Fox Mulder Transform

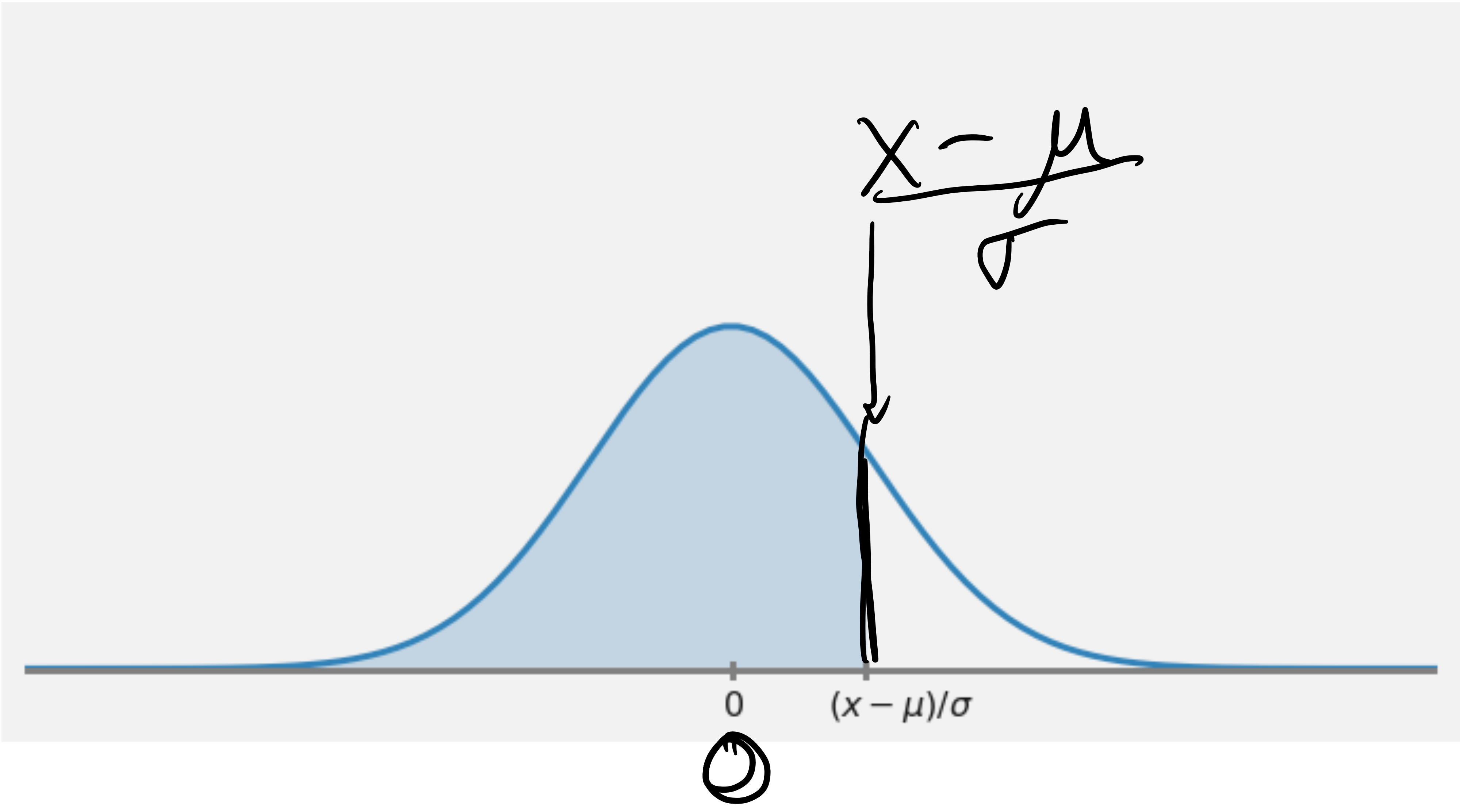
Nonstandard normals

- Normal distributions that are not standard can be turned into standard normals so, so, so easily!
- **Proposition**: if X is a normal distribution with mean mu and standard deviation sigma, then Z is a standard normal distribution if:

$$Z = \frac{X - \mu}{\sigma} \quad \text{or} \quad X = \sigma Z + \mu$$



$\Leftrightarrow$



Area Intuition

https://www.intmath.com/counting-probability/normal-distribution-graph-interactive.php

Standard normals in action

- **Example**: The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions.
- The article Fast-Rise Brake Lamp as a Collision-Prevention Device* suggests that **reaction time for an in-traffic response to a brake signal** from standard brake lights can be modeled with a normal distribution having **mean value** $\underbrace{\textbf{1.25 sec}}_{\mu}$ and **standard deviation of** $\underbrace{\textbf{.46 sec}}_{\sigma}$.
- What is the probability that reaction time is between 1.00 sec and 1.75 sec?

$$P(1.00 \le X \le 1.75) = P\left(\frac{1.00-1.25}{0.46} \le \frac{X-1.25}{0.46} \le \frac{1.75-1.25}{0.46}\right)$$

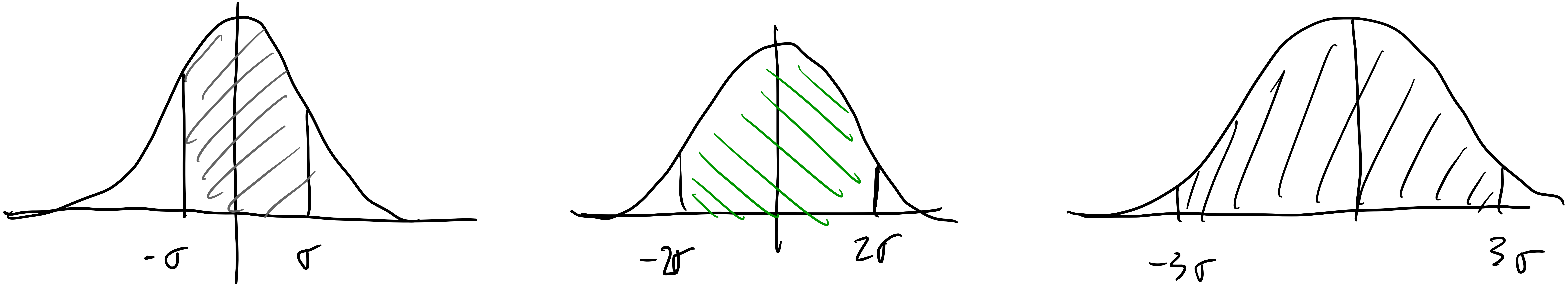
$$= P(-0.543 \le Z \le 1.09)$$

$$= \Phi(1.09) - \Phi(-0.543) = \boxed{0.568}$$

* (Ergonomics, 1993: 391–395)

Standard Normals in action

- So what are some common things that come up with normals?



0.683

1 in 3.2

0.954

1 in 22

0.997

1 in 370.4