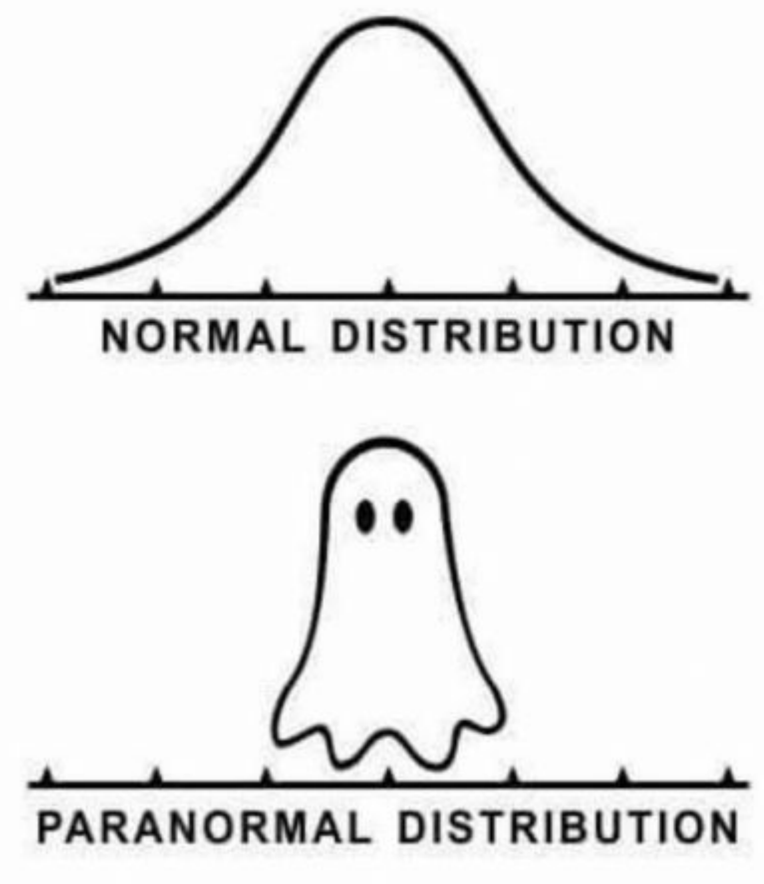
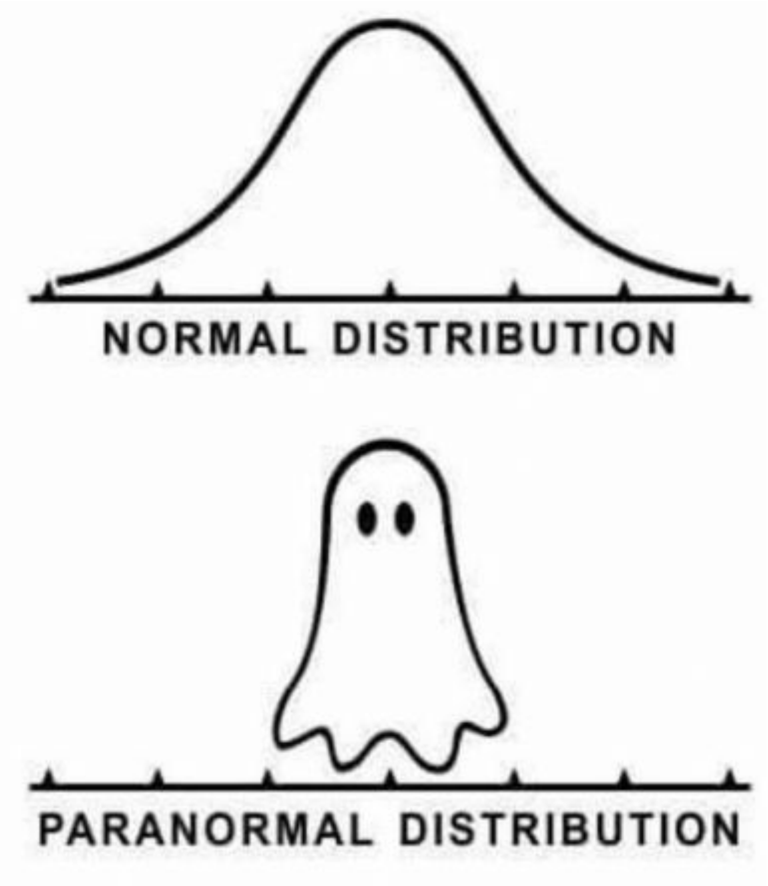
Lecture 12: The Normal Distribution



Announcements and reminders

- Practicum 1 due Monday, 11:59 PM



Previously, on CSCI 3022…

Definition: A random variable X is **continuous** if for some function $f : \mathbb{R} \to \mathbb{R}$ and for any numbers a and b with $a \leq b$,

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

The function f must satisfy:

1) $f(x) \geq 0$ for all x, and 2) $\int_{-\infty}^{\infty} f(x)\ dx = 1$

Definition: The **cumulative distribution** (or density) **function** of X is defined such that

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

The Normal distribution

The **normal distribution** (AKA, Gaussian distribution) is probably the most important and widely used distribution in probability and statistics.

Many populations have distributions well-approximated by a normal distribution.

It's **very important** to check that Normal is a good approximation though! And **justify**.

Examples:

- Height, Weight, Other physical attributes
- Scores on a test
- Time it takes to travel

Consider: Why might Normal be an issue?

The Normal distribution

Definition: A continuous random variable X has a **normal (or Gaussian) distribution** with parameters μ and σ^2 if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

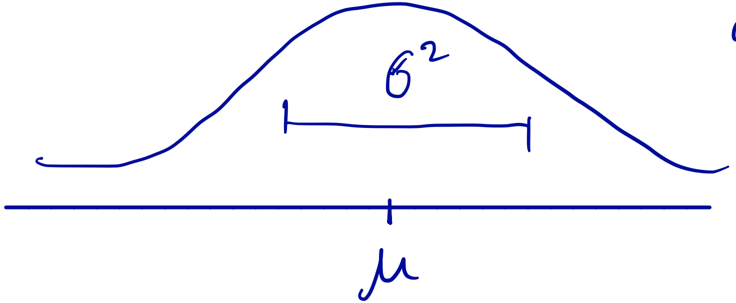
standard deviation

np.exp(...)

We say $X \sim N(\mu, \sigma^2)$

If we didn't have this σ, then it wouldn't be normalized $\left(\int_{-\infty}^{\infty} f(x)\,dx = 1\right)$

Let's play around with this distribution: https://academo.org/demos/gaussian-distribution/



except a real normal is symmetric

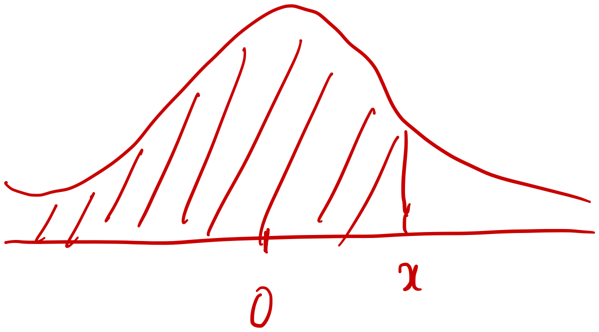
The Standard Normal distribution

Definition: The normal distribution with parameter values $\mu = 0$ and $\sigma^2 = 1$ is called the <u>**standard normal distribution**</u>.

Question: What is the pdf of the standard normal distribution? $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

cdf : $F(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$

$\int e^{-x^2} dx \leftarrow$ no closed form!



The Standard Normal distribution

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

Question: What is the pdf of the standard normal distribution?

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

The Standard Normal distribution

Definition: The normal distribution with parameter values $\mu = 0$ and $\sigma^2 = 1$ is called the <u>**standard normal distribution**</u>.

Question: What is the pdf of the standard normal distribution?

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

A standard normal random variable is usually denoted Z

Recall: The normal distribution does not have a closed form cumulative distribution function

→ We use special notation to denote the cdf of the **standard** normal distribution:

$$\boxed{\Phi(z) = P(Z \leq z)} = F(z) = \int_{-\infty}^{z} f(x)\,dx$$

→ And usually we just look up values for $\Phi(z)$ in a table

The Standard Normal distribution

$\Phi(z) = \texttt{scipy.stats.norm.cdf}(z)$

↳ import scipy.stats as stats

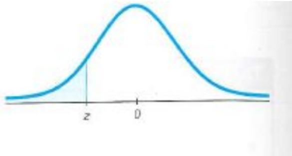
The standard normal dist. **rarely** occurs in real life.

Instead, we take non-standard normal distributions, and **standardize** them using a simple transformation.

Recall: For computing probabilities, having a cdf is just as good (or better!) as having a pdf

Back in MY day you had to look up values of the standard normal cdf in **tables** in the back of textbooks.

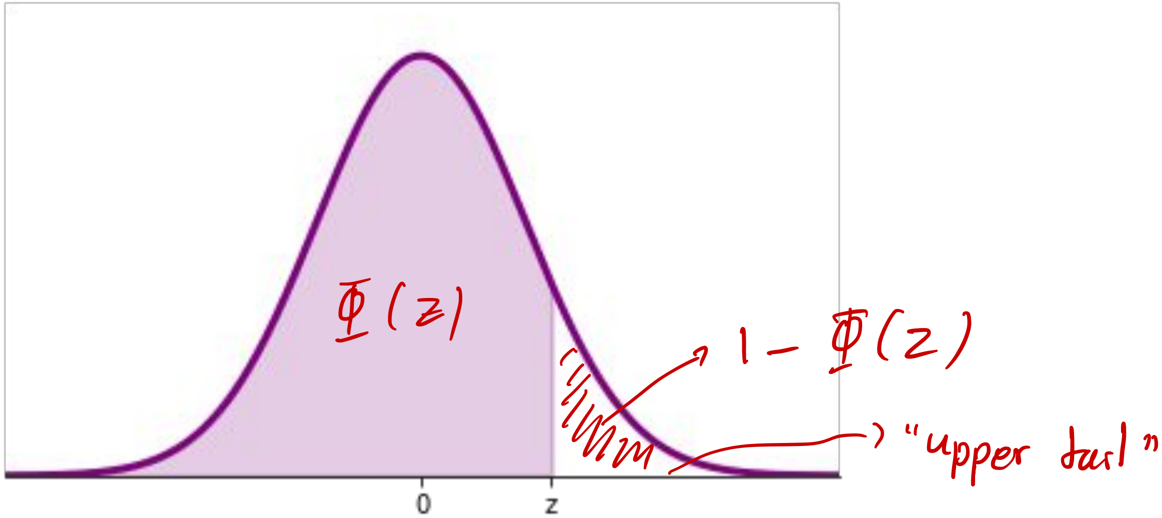
NEGATIVE *z* Scores



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007

The Standard Normal distribution

$\Phi(z)$ = shaded area

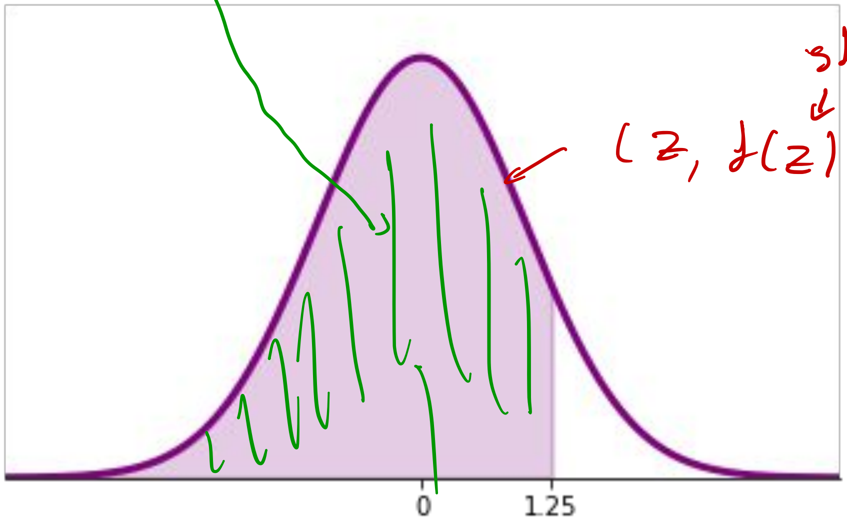


The Standard Normal distribution

Example: What is P(Z ≤ 1.25) ? = stats.norm.cdf(1.25)

stats.norm.pdf(z)

(z, f(z))



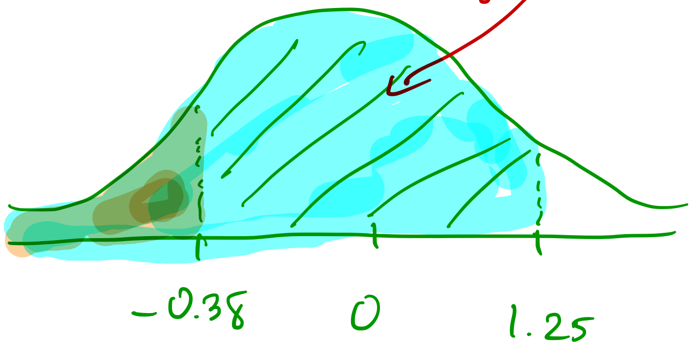
The Standard Normal distribution

$P(2 \le X \le 4) = \sum_{a_i=2}^{4} P(X = a_i)$

roll fair dice b/w 2 to 4

$P(-0.38 \le Z \le 1.25) = \int_{-0.38}^{1.25} f(x)\,dx$

Example: What is P(Z ≥ 1.25) ?



Example: What is P(Z ≤ -1.25) ?

Example: How can we calculate P(-0.38 ≤ Z ≤ 1.25) ?

$\Phi(1.25) = \int_{-\infty}^{1.25} f(x)\,dx$

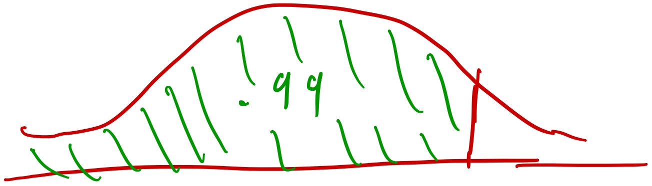
$\Phi(0.38) = \int_{-\infty}^{-0.38} f(x)\,dx$

$P(-0.38 \le Z \le 1.25) = \Phi(1.25) - \Phi(-0.38)$

Flip it *and* Reverse it

$Q_3 = 75^{th}$ percentile

Example: What is the 99th percentile of N(0, 1)?



$z = ?$

We have tables that tell us **area**… but we were given the area.

This is the **inverse** problem to P(Z ≤ z) = 0.99 $= \Phi(z)$

?

$F(z) = 0.99$

$F^{-1}(.99) = z$

What about in Python?

- scipy.stats.norm.cdf $(F(x))$
- scipy.stats.norm.pdf $(f(x))$
- scipy.stats.norm.ppf $F^{-1}(x)$

↳ percent point function

- scipy.stats.rvs(....) random samples

Example:

if $F(Q_3) = 0.75$,

then 1) stats.norms.cdf$(Q_3) = .75$

⋮

2) stats.norm.ppf$(.75) = Q_3$

Flip it *and* Reverse it

Example: What is the 99th percentile of N(0, 1)?

We have tables that tell us **area**… but we were given the area.

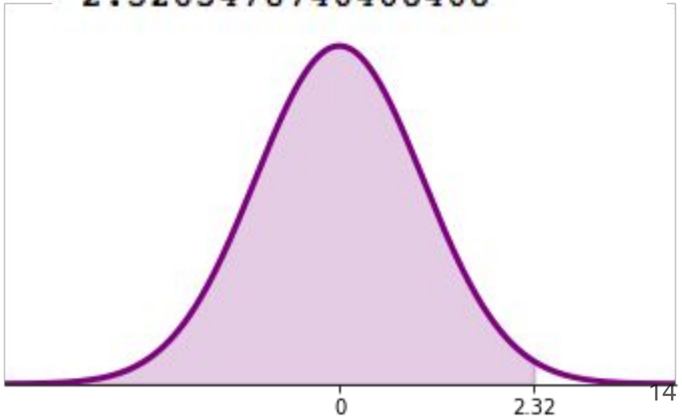
This is the **inverse** problem to $P(Z \leq z) = 0.99$

What about in Python?

- scipy.stats.norm.cdf
- scipy.stats.norm.pdf
- scipy.stats.norm.ppf

```python
from scipy import stats
stats.norm.ppf(.99)
```

```
2.3263478740408408
```

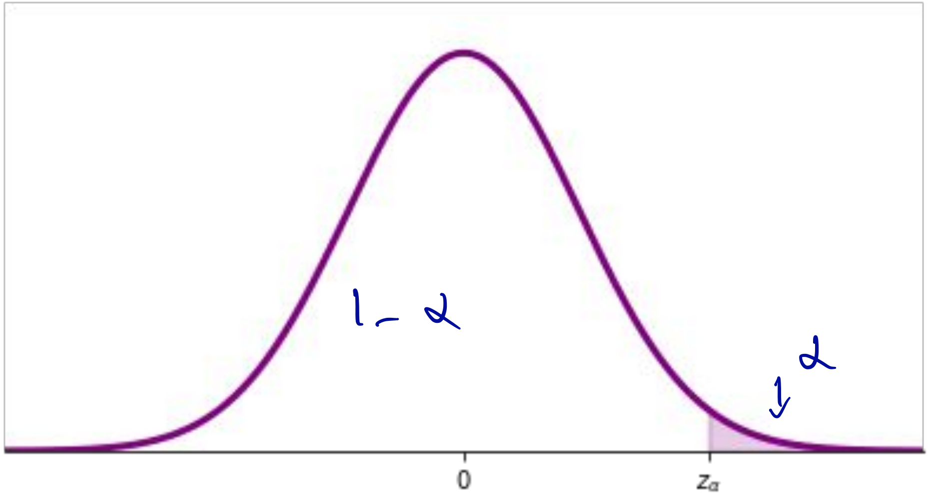


The Critical Value

Notation: We say z_α is the **critical 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 of the area under the curve lies to the **right** of z_α

$F(z_\alpha) = 1 - \alpha \iff z_\alpha = \text{ppf}(1-\alpha)$

Note that other books/resources might use different conventions.
Be careful and **use sanity checks**!

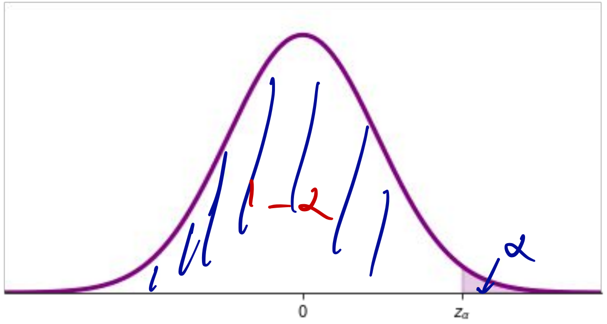


The Critical Value

Notation: We say z_α is the **critical 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 of the area under the curve lies to the **right** of z_α

Question: What is the relationship between z_α and the cdf?

$$\int_{-\infty}^{\infty} f(z)\,dz, \quad \Phi(z_\alpha) = 1-\alpha = P(Z \le z_\alpha)$$



Question: What is the relationship between z_α and percentiles?

z_α is the $100(1-\alpha)^{th}$ percentile

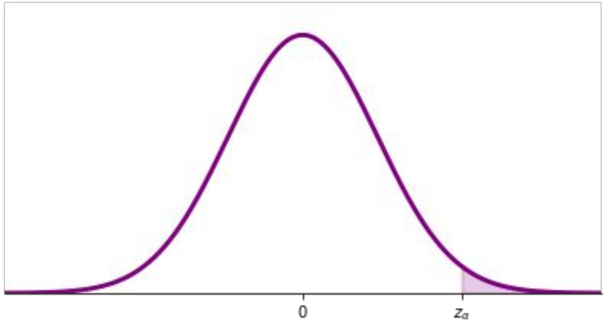
Ex: $\alpha = 0.5 \rightarrow 100 \times (1-0.5) = 50^{th}$ perc. (median) ∴ $z_{0.50}$ is the median)

The Critical Value

Notation: We say z_α is the **critical 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 of the area under the curve lies to the **right** of z_α

Question: What is the relationship between z_α and the cdf?

$$P(Z \geq z_\alpha) = \alpha = 1 - P(Z \leq z_\alpha) = 1 - \Phi(z_\alpha)$$



Question: What is the relationship between z_α and percentiles?

z_α is the $100(1-\alpha)^{th}$ percentile

Non-standard Normal Distributions

$X \sim N(\mu, \sigma^2)$

Non-standard normal distributions can be turned into standard normals really easily

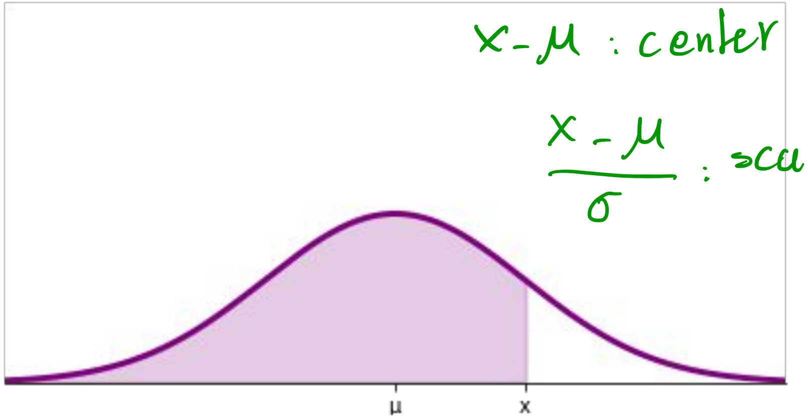
Proposition: If X is a normally distributed random variable with mean μ and standard deviation σ, then Z follows a standard normal distribution if we define:

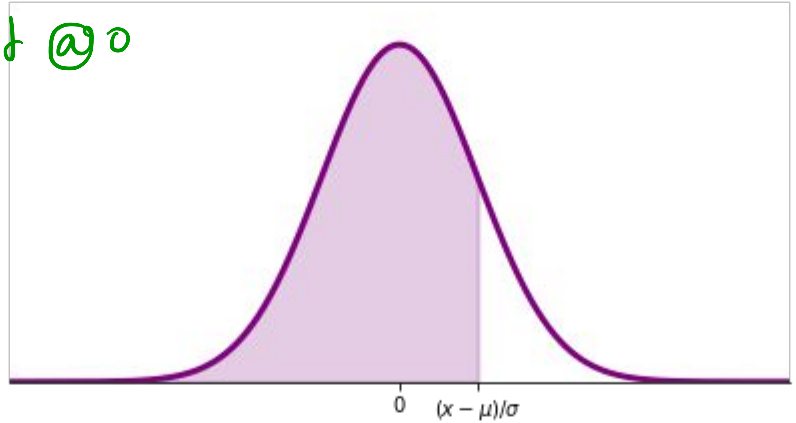
$$\boxed{Z = \frac{X - \mu}{\sigma}} \quad \text{and} \quad \boxed{X = \sigma Z + \mu}$$

Box-Muller transforms

$x - \mu$: center the dist @ 0

$\frac{x - \mu}{\sigma}$: scale





Brake lights!

Example: The time it takes a driver to react to brake lights on a decelerating vehicle is important for not getting into rear-end collisions.

The article "Fast-Rise Brake Lamp as a Collision Prevention Device" (linked here) suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled as a normal distribution having mean value 1.25 s and standard deviation 0.46 s.

Question: What is the probability that a reaction time is between 1.0 s and 1.75 s?

$P(1.0 \le x \le 1.75)$

standardize! Box-Moller $= z = \dfrac{x - 1.25}{0.46}$

$$P(1.0 \le x \le 1.75) = P\left(\frac{1.0 - 1.25}{0.46} \le z \le \frac{1.75 - 1.25}{0.46}\right)$$

$$= P\left(-\frac{0.25}{0.46} \le z \le \frac{.5}{0.45}\right) = \boxed{\Phi(1.09) - \Phi(-.54)}$$

$\Rightarrow$ stats.norm.cdf(1.75, loc = 1.25, scale=0.46) − stats.norm.cdf(- -)

= norm.cdf(1.09) − norm.cdf(−.54)

Brake lights!

Example: The time it takes a driver to react to brake lights on a decelerating vehicle is important for not getting into rear-end collisions.

The article “Fast-Rise Brake Lamp as a Collision Prevention Device” (linked here) suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled as a normal distribution having mean value 1.25 s and standard deviation 0.46 s.

Question: What is the probability that a reaction time is between 1.0 s and 1.75 s?

Follow-up question: What might be a potential problem with using a normal distribution? How can we check if it is much of an issue?

What just happened?

- We learned about the normal distribution!
- And the **standard normal distribution**
- And how to take any ol' normal random variable and **standardize it**