Introduction to Data Science With Probability and Statistics

Lecture 15: The Normal Distribution

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

What will we learn today?

- ❑ Normal Distribution
- ❑ Standard Normal Distribution
- ❑ Critical values

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 5, Section 5*

Review from Last Time

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

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

b

The function must satisfy:

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

The **cumulative distribution 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 the Gaussian distribution) is probably the most important and widely used distributions in probability and statistics.

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

It is very important to check that Normal is a good approximation. And to justify!

Examples:

- Height, weight
- Scores on a test
- Time it takes to travel

Normal Distribution

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

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

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

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

μ : mean

σ : standard deviation.

The Standard Normal Distribution

When $\mu = 0$ and $\sigma^2 = 1$, the normal distribution is called the **standard normal distribution**.

- What is the pdf of the standard normal distribution?

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

- What is the cdf of the standard normal distribution?

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \underbrace{\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,dt}_{\text{numerical approximation is used.}}$$

The Standard Normal Distribution

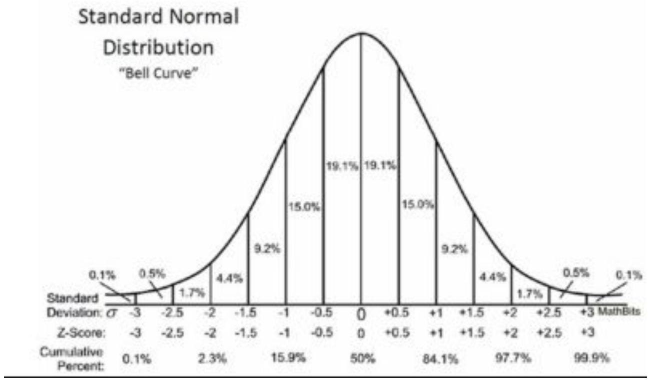
A standard normal random variable is usually denoted Z.

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**:

$$\Phi(z) = P(Z \leq z)$$

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



import scipy.stats as stats

$\Phi(z) = stats.norm.cdf(z)$

The Standard Normal Distribution

The standard normal distribution rarely occurs in real life.

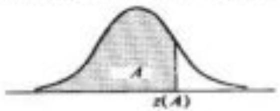
Instead, we take non-standard normal distribution, and standardize them using a simple transformation.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Normal Distribution

Table C-1. Cumulative Probabilities of the Standard Normal Distribution.

Entry is area A under the standard normal curve from $-\infty$ to $z(A)$



$$P(Z \le 1.51) = \Phi(1.51)$$

$$= \underline{0.9345}$$

Source

The Standard Normal Distribution

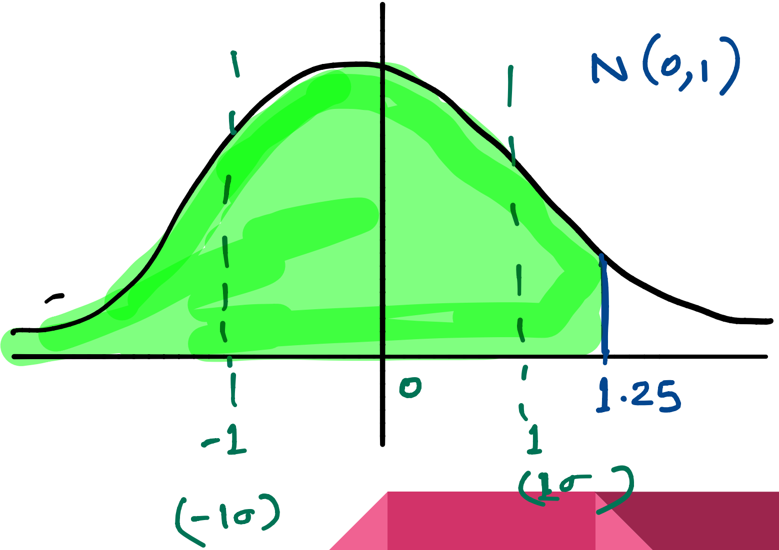
Example: What is $P(Z \leq 1.25)$?

$$P(Z \leq 1.25) = \Phi(1.25)$$

use either table OR.

python :

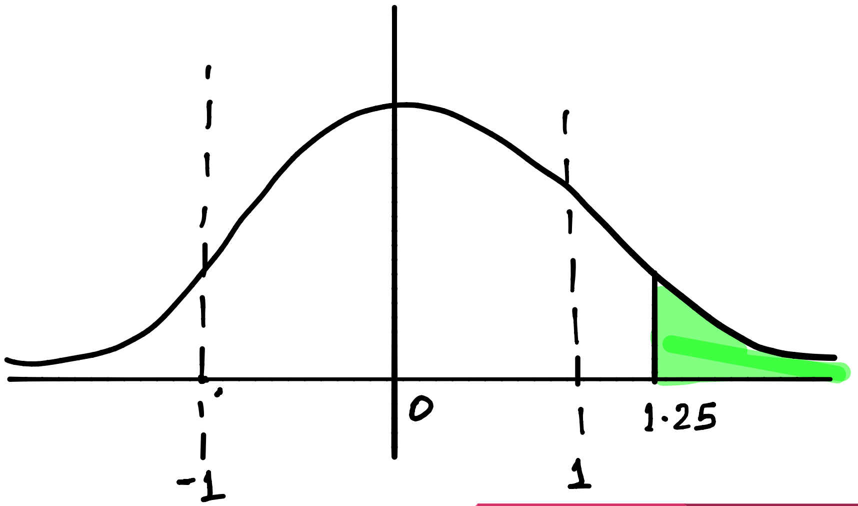
$$\boxed{\begin{array}{l} \text{stats.norm.cdf}(1.25) \\ = 0.8944 \end{array}}$$



The Standard Normal Distribution

Example: What is $P(Z \geq 1.25)$?

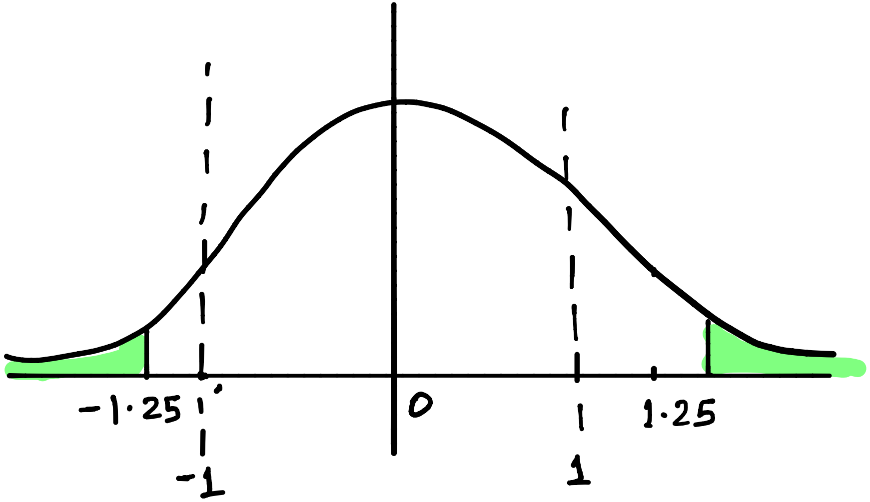
$$P(Z \geq 1.25) = 1 - P(Z < 1.25)$$
$$= 1 - \phi(1.25)$$
$$= 0.1056$$



The Standard Normal Distribution

Example: What is $P(Z \leq -1.25)$?

$$P(Z \leq -1.25) = P(Z \geq 1.25)$$
$$= 0.1056$$



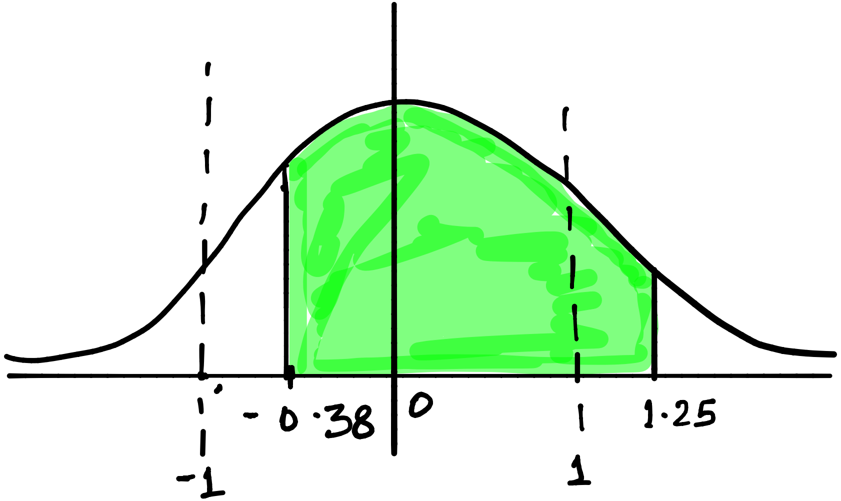
The Standard Normal Distribution

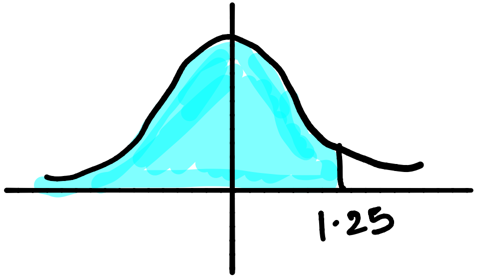
Example: What is $P(-0.38 \le Z \le 1.25)$?

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

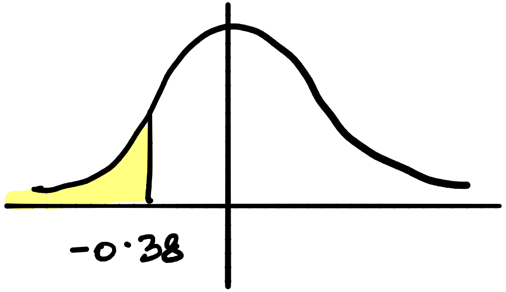
$$= \phi(1.25) - \phi(-0.38)$$

$$= \text{stats.norm.cdf}(1.25) - \text{stats.norm.cdf}(-0.38)$$





$-$

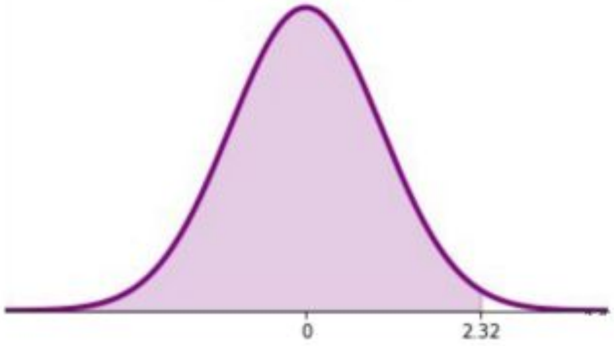


The Standard Normal Distribution

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

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

2.3263478740408408



In Python:

- ϕ scipy.stats.norm.cdf
- f scipy.stats.norm.pdf
- ϕ^{-1} scipy.stats.norm.ppf

stats.norm.cdf(2.32) = 0.99

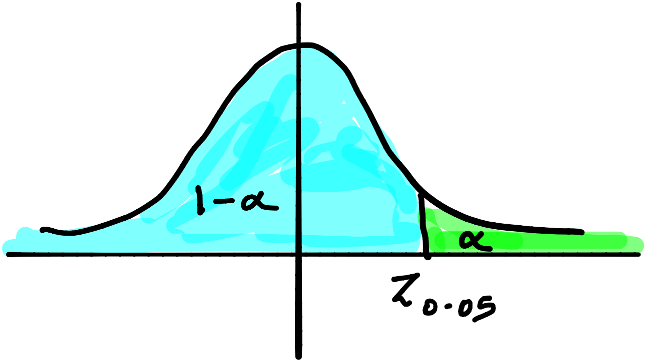
stats.norm.ppf(0.5) = 0

The Critical Value

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

- Note that other books might use different conventions, so be careful and sanity check often.

if $\alpha = 0.05$ then $z_{0.05}$ is the critical value.



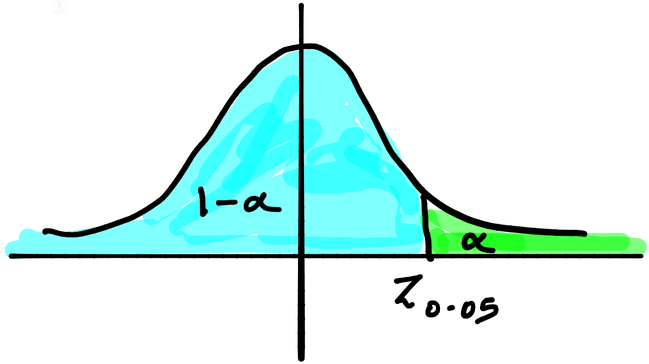
What is the relationship between z_α and the cdf?

What is the relationship between z_α and percentiles?

The Critical Value

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

- Note that other books might use different conventions, so be careful and sanity check often.



What is the relationship between z_α and the cdf?

$$\phi(z_{0.05}) = 1-\alpha$$

What is the relationship between z_α and percentiles?

Z_α is the $100(1-\alpha)$th percentile

eg: $\alpha = 0.05$; $Z_{0.05}$ is the 95% percentile value

Non-Standard Normal Distributions

Non-standard normal distributions can be turned into standard normal.

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:

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

std
Normal

Normal.

Brake Lights

Example: The time it takes a driver to react to brake lights on a decelerating vehicle is important to understand. 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 as a normal distribution having a mean value of 1.25 seconds and standard deviation 0.46 seconds.

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

Brake Lights

Example: The time it takes a driver to react to brake lights on a decelerating vehicle is important to understand. 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 as a normal distribution having a mean value of 1.25 seconds and standard deviation 0.46 seconds.

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

$$R \sim N(1.25, 0.46^2)$$

$$P(1.0 \leq R \leq 1.75)$$

$$= P(1.0 - \mu \leq R - \mu \leq 1.75 - \mu)$$

$$= P\left(\frac{1.0 - \mu}{\sigma} \leq \frac{R - \mu}{\sigma} \leq \frac{1.75 - \mu}{\sigma}\right)$$

we want $R \sim N(1.25, 0.46^2)$ to transform to $Z \sim N(0,1)$

Brake Lights - continued

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

$$= P(-0.54 \le z \le 1.09)$$

$$= \phi(1.09) - \phi(-0.54)$$

$$\approx 0.5674$$

Next Time:

- Central Limit Theorem !