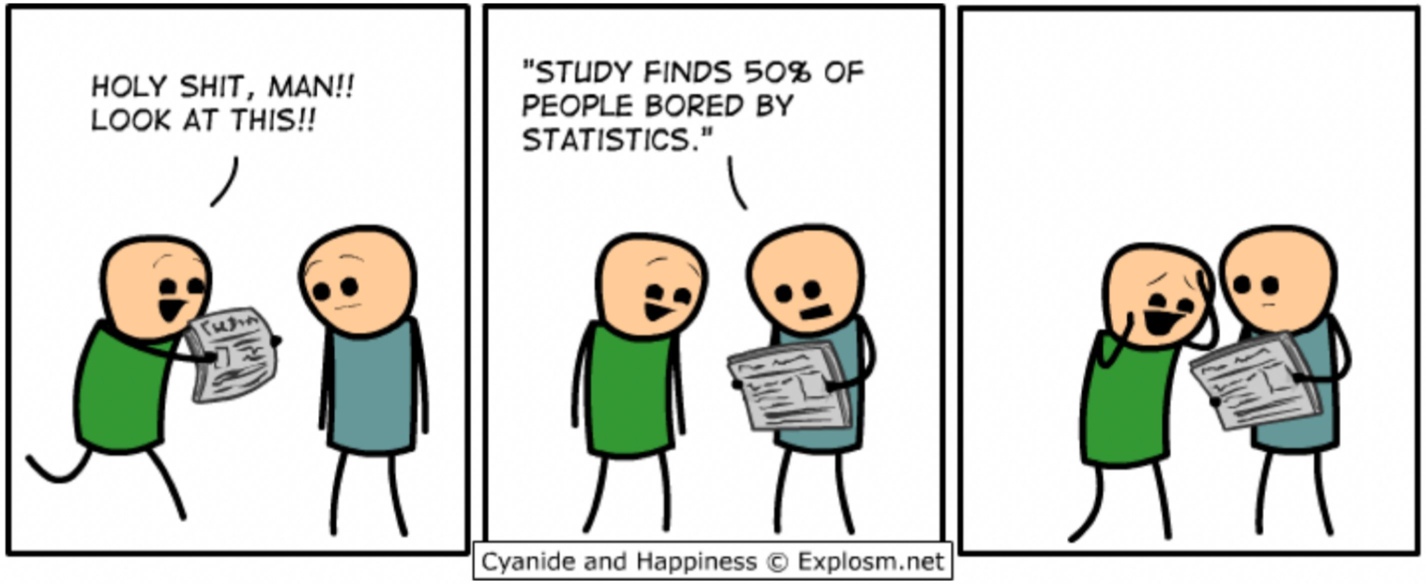
CSCI 3022: Intro to Data Science
Lecture 13: The Normal Distribution

Rachel Cox

Department of Computer Science



Announcements & Reminders

Practicum I – Due tonight at 11:59 pm • Late Days do not apply

Quizlet 04 – Due Wednesday at 11:59 pm on Canvas

Quizlet 05 – Due Wednesday March 11^{th}

Midterm Exam

- Average ~ 77%
- Solutions posted
- Regrade requests – submit through Gradescope.
  - Due by Monday March 11^{th}

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 $a \leq b$,

b

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

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

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

μ – mean (for the population, not a sample)

σ^2 – variance

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

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

What about the cdf of the Normal dist?

$$P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

No closed form.

$\int e^{-x}\,dx = -e^{-x} + C$

$\int e^{-x^2}\,dx = ?$ $\quad \mathrm{erf}(x)$

The Standard Normal Distribution

$e^{-\frac{1}{2}x^2}$ – use the Taylor series expansion to approximate

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?

pdf of normal: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ let $\sigma = 1$, $\mu = 0$ $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x)^2}$

- What is the cdf of the standard normal distribution?

$$F(x) = \int_{-\infty}^{x} f(x) = \underbrace{\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx}_{\text{no closed form!}}$$

→ use python to approximate.

The Standard Normal Distribution

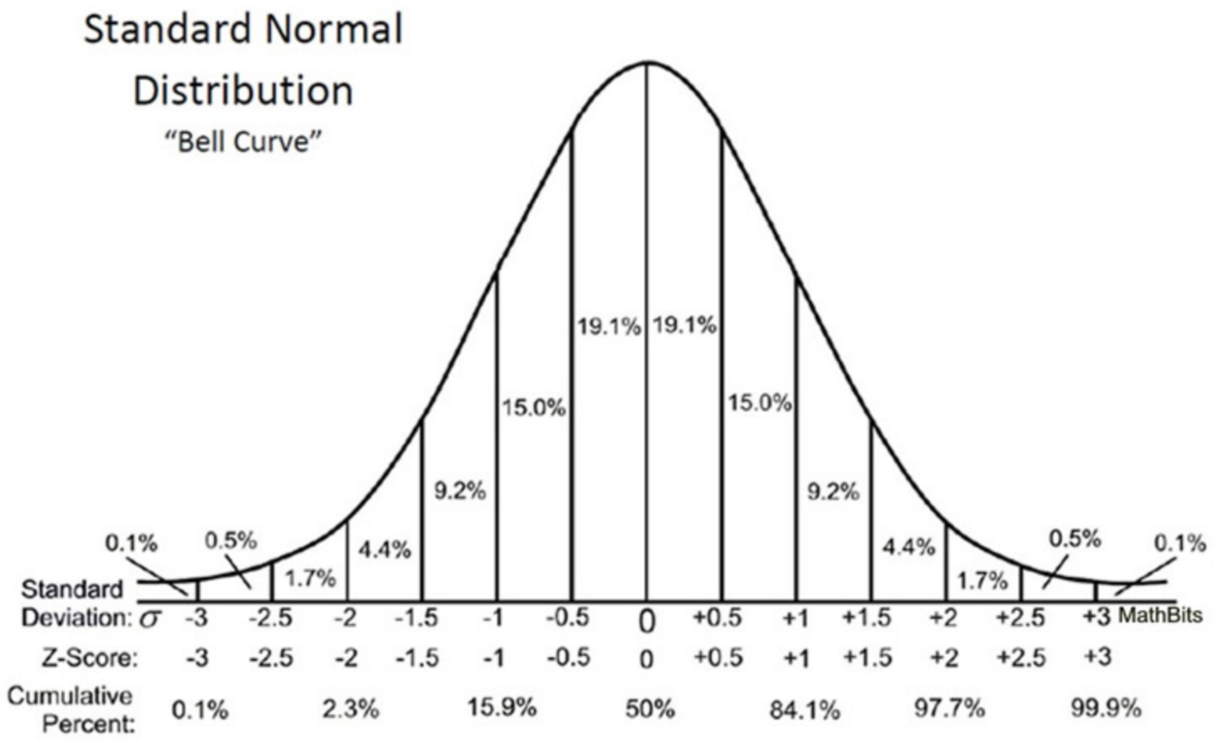
A standard normal random variable is usually denoted Z.

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.



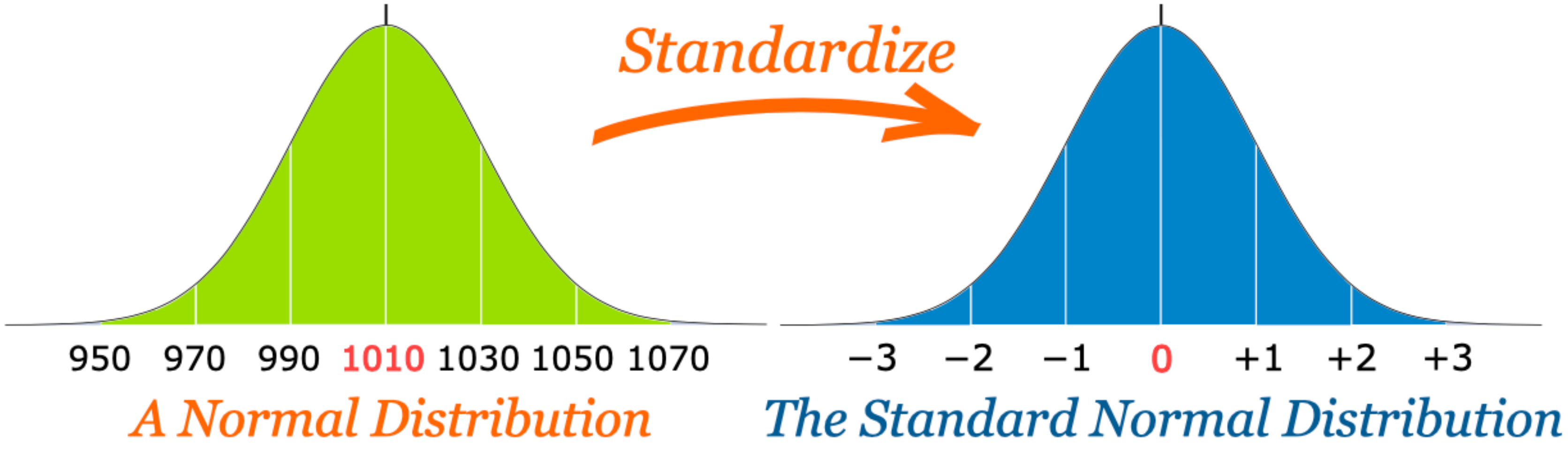
import scipy.stats as stats

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

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.



pdf f

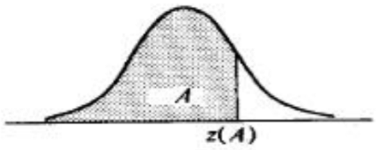
cdf F

Standard normal cdf: Φ

Normal Distribution

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

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



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

$$P(Z \le 1.33) = \Phi(1.33)$$

$$= \int_{-\infty}^{1.33} \frac{1}{\sqrt{2\pi}} e^{-x^2} dx$$

$$= .9082$$

Source

in python: stats.norm.cdf(1.33)

The Standard Normal Distribution

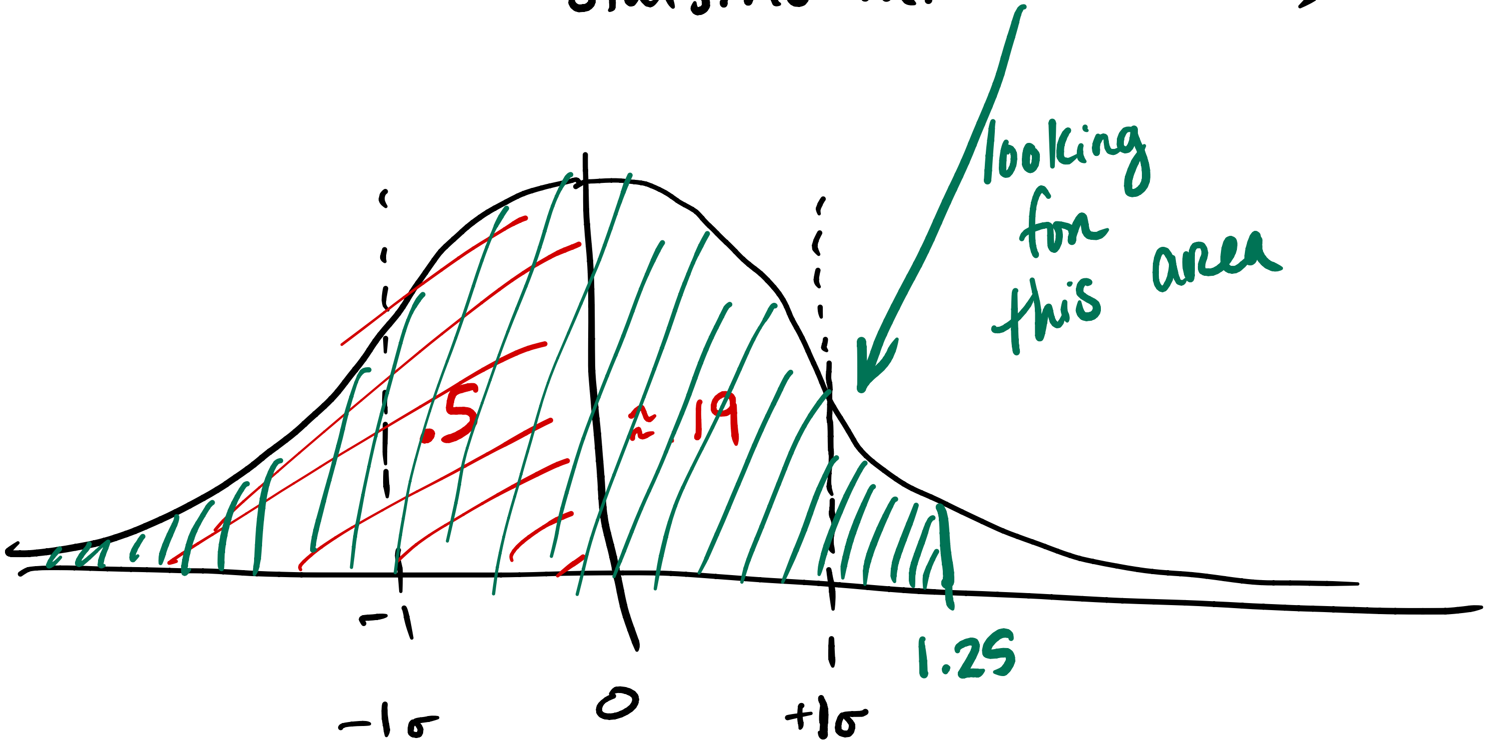
Z – indication that we have a Standard normal Random variable

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

Φ – cdf for standard normal

$P(Z \leq 1.25)$

$= \text{stats.norm.cdf}(1.25) \approx .8944$



The Standard Normal Distribution

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

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

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

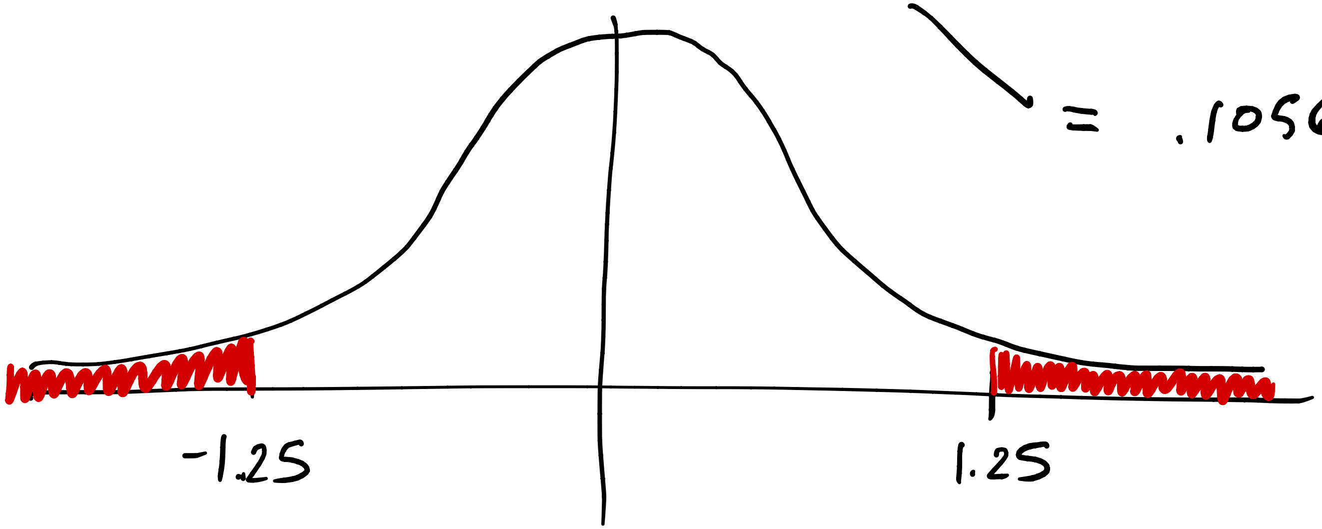
$$= 1 - 0.8944$$

$$\approx .105649$$

The Standard Normal Distribution

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

$P(Z \geq 1.25) \approx .1056$



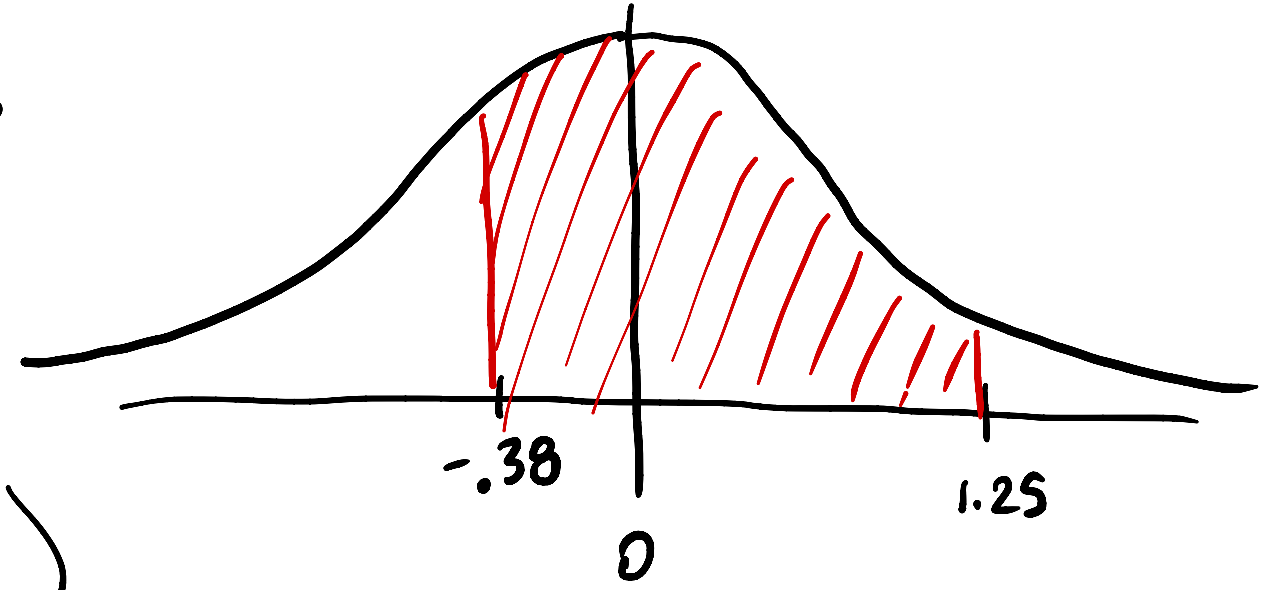
$= .1056$

The Standard Normal Distribution

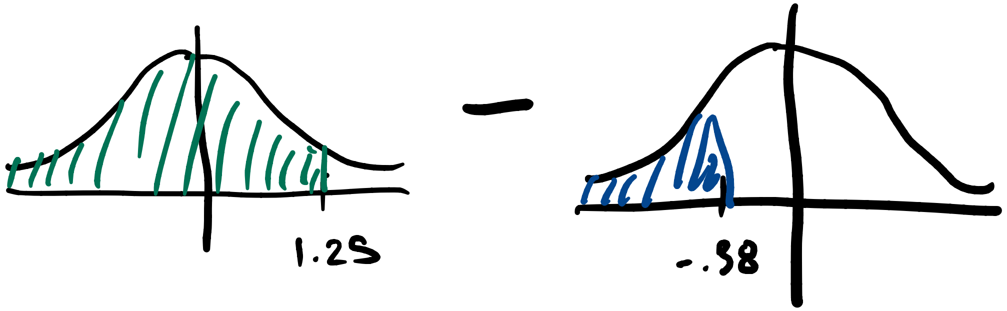
import scipy.stats as stats

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

$P(-.38 \le Z \le 1.25)$



$= \Phi(1.25) - \Phi(-.38)$



= stats.norm.cdf(1.25) − stats.norm.cdf(−.38)

The Standard Normal Distribution

In Python:

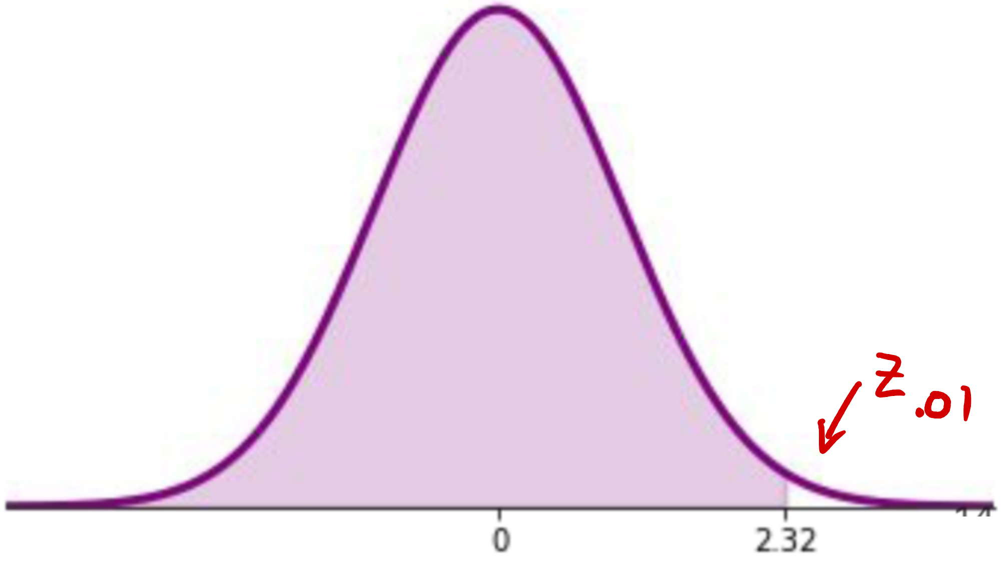
$\Phi(x)$ • scipy.stats.norm.cdf
$f(x)$ • scipy.stats.norm.pdf
$\Phi^{-1}(x)$ • scipy.stats.norm.ppf

"percent point funct"

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

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

```
2.3263478740408408
```



e.g

stats.norm.cdf(2.32) = .99

Stats.norm.ppf(.99) = 2.32

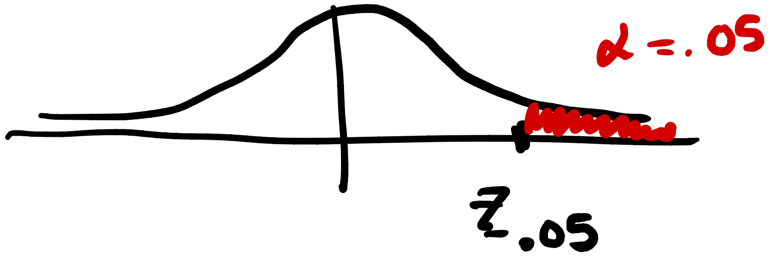
stats.norm.ppf(.5) = 0

The Critical Value

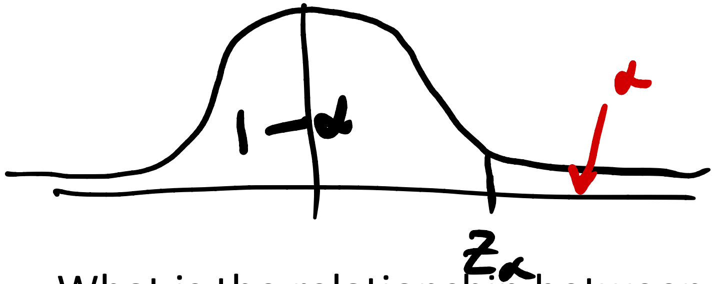
We say z_α is the **critical value** of Z under the standard normal distribution that gives a certain tail area. In particular, it is the Z value such that exactly α of the area under the curve lies to the **right** of z_α

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

If $\alpha = .05$ $\quad z_{.05}$



What is the relationship between z_α and the cdf?



$$\boxed{\begin{aligned}\Phi(z_\alpha) &= 1-\alpha \\ \alpha &= 1-\Phi(z_\alpha)\end{aligned}}$$

What is the relationship between z_α and percentiles?

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

e.g. $\alpha = 0.1$ $\quad z_{0.1}$ is the $100(1-0.1)^{th}$ $\boxed{90^{th} \text{ percentile}}$

Non–Standard Normal Distributions

Remember HW!

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:

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

Z - standard normal

X - 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.

Reaction time $N(1.25, 0.46^2)$

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

use the transformation

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

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

$$= P(1.0-\mu \le X-\mu \le 1.75-\mu)$$

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

$$= P\left(\frac{1.0-1.25}{0.46} \le Z \le \frac{1.75-1.25}{0.46}\right) = P(-.54 \le Z \le 1.09)$$

Brake Lights – continued

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

So $P(-.54 \le X \le 1.09) = \Phi(1.09) - \Phi(-.54)$

$= \text{stats.norm.cdf}(1.09)$

$- \text{stats.norm.cdf}(-.54)$

≈ 0.5675449

Notebook 11
Handwritten Examples

Exercise 2 : pmf for X :

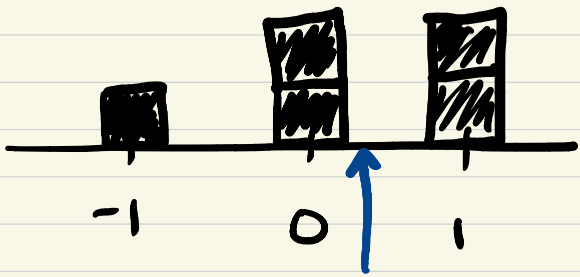
Part A:

$$f(x) = \begin{cases} \frac{1}{5} & X = -1 \\ \frac{2}{5} & x = 0 \\ \frac{2}{5} & x = +1 \end{cases}$$

$$E[X] = -1 \cdot \tfrac{1}{5} + 0 \cdot \tfrac{2}{5} + 1 \cdot \tfrac{2}{5}$$

$$= -\frac{1}{5} + \frac{2}{5}$$

$$= \frac{1}{5}$$



Part B : Suppose we have a R.V.

$Y = X^2$

What is the pmf for Y ?

$P(X=-1) = \frac{1}{5}$ | $P(X=1) = \frac{2}{5}$

when $x = -1$, $Y = (-1)^2 = 1$ | $X = 1 \Rightarrow Y = 1$

$P(X=0) = \frac{2}{5}$

when $x=0 \Rightarrow Y = 0^2 = 0$

$\frac{2}{5} = P(X=0) = P(Y=0)$

$P(Y=1) = P(X=-1) + P(X=1)$

$= \frac{1}{5} + \frac{2}{5}$

$= 3/5$

$P(Y=0) = 2/5$

$P(Y=1) = 3/5$

Part C $\quad E[Y] = 0 \cdot \frac{2}{5} + 1 \cdot 3/5$

$= 3/5$

compare

* $E[X^2] = (-1)^2 \cdot \frac{1}{5} + 0^2 \cdot \frac{2}{5} + 1^2 \cdot \frac{2}{5}$

$= 3/5$

Part D $Var(X)$

$$Var(x) = E[x^2] - \left(E[x]\right)^2$$

$$= \frac{3}{5} - \left(\frac{1}{5}\right)^2$$

$$= \frac{3}{5} - \frac{1}{25}$$

$$= 14/25$$

Exercise 3 $E[X] = 4$.

$$Var(X) = 2$$

Part A : Compute $E[X^2]$

$$\boxed{Var(X) = E[X^2] - (E[X])^2}$$

$$E[X^2] = Var(X) + (E[X])^2$$

$$= 2 + 4^2$$

$$= 18 .$$

Part B : Expectation + Variance of $Y = 1 - 2X$

$$E[Y] = E[1-2X]$$

$$= E[1] - 2E[X]$$

$$= 1 - 2 \cdot 4$$

$$= -7$$

$E[Y] = -7$

$\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2$

$Y = 1 - 2X$

$Y^2 = (1-2X)^2$

$= 1 - 4x + 4x^2$

$E[Y^2] = E[1 - 4x + 4x^2]$

$= 1 - 4E[x] + 4E[x^2]$

$= 1 - 4 \cdot 4 \quad + 4 \cdot 18$

$= 57$

$\boxed{\mathrm{Var}(Y) = 57 - (-7)^2 = 8}$

Next Time:

- ❖ Central Limit Theorem