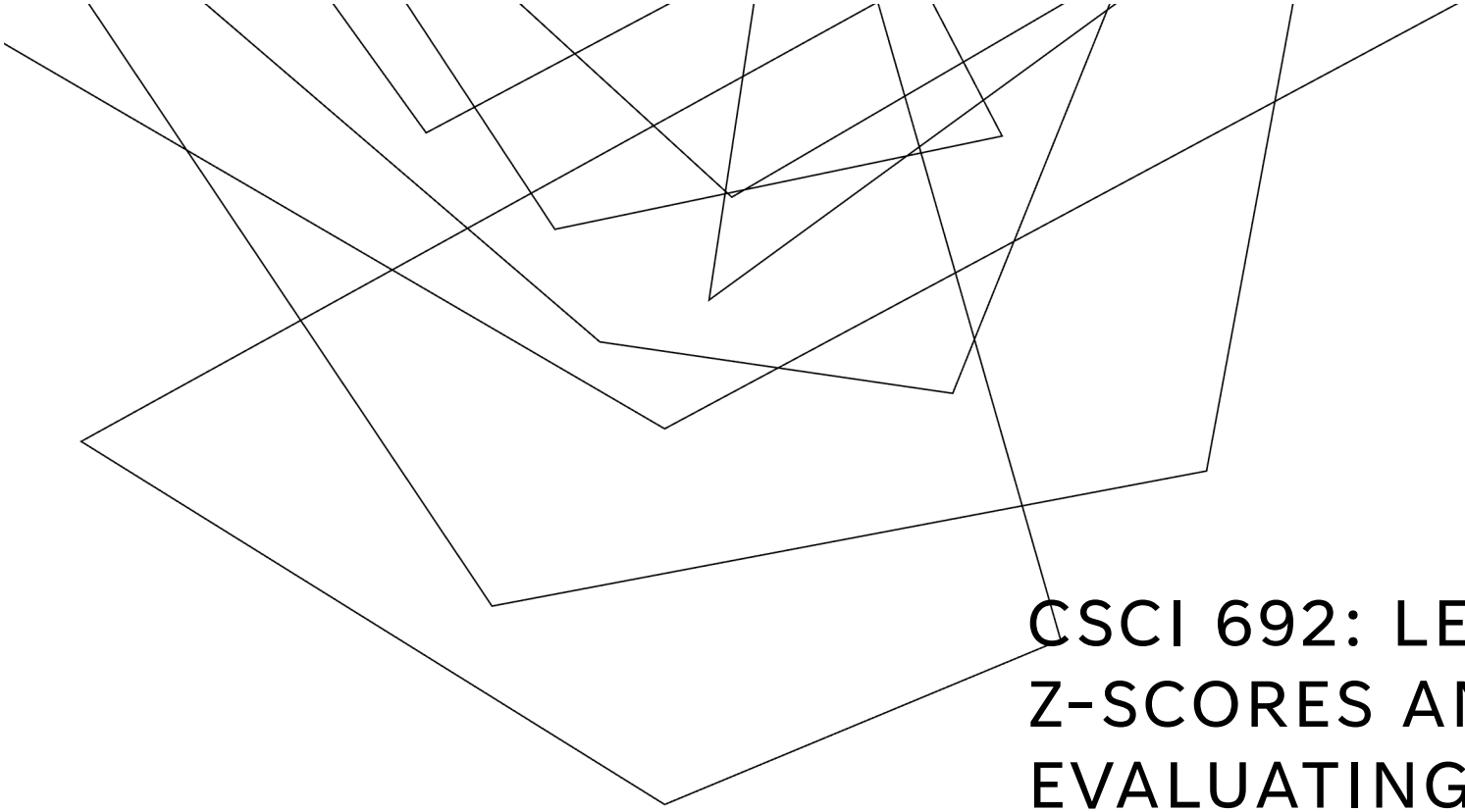


CSCI 692: LECTURE 8 Z-SCORES AND EVALUATING GAUSSIANS

Professor David Harrison

OFFICE HOURS

Tuesday	4:00-5:00 PM
Wednesday	12:30-2:30 PM

HOMEWORK 3

Handed out Thursday, Feb 21.

Today is February 15

NOTE REGARDING EXAMS

Note in homework 2:

Note regarding the midterm and final: The midterm and final will be written, so students will not have access to Databricks or Jupyter or Python. The questions asked here on an exam would be computed on small datasets as are used for question 31 in Part 7 and all the problems in Parts 8 and 9. I recommend that you answer the questions in these sections without using Python or a calculator. The problems are not difficult, and doing them by hand may prepare you for answering such questions on the exams.

DATES OF INTEREST

February 8	HW2 handed out
February 15	HW2 due,
February 15 /16->22	HW3 handed out
February 22->26	HW3 due
February 27	Review
February 29	Midterm (must be before progress reports)
March 4	Progress Reports
March 8	Deadline for Withdrawal
March 9-17	Spring Break

BLACKBOARD & GITHUB

Slides up through lecture 7 on blackboard.

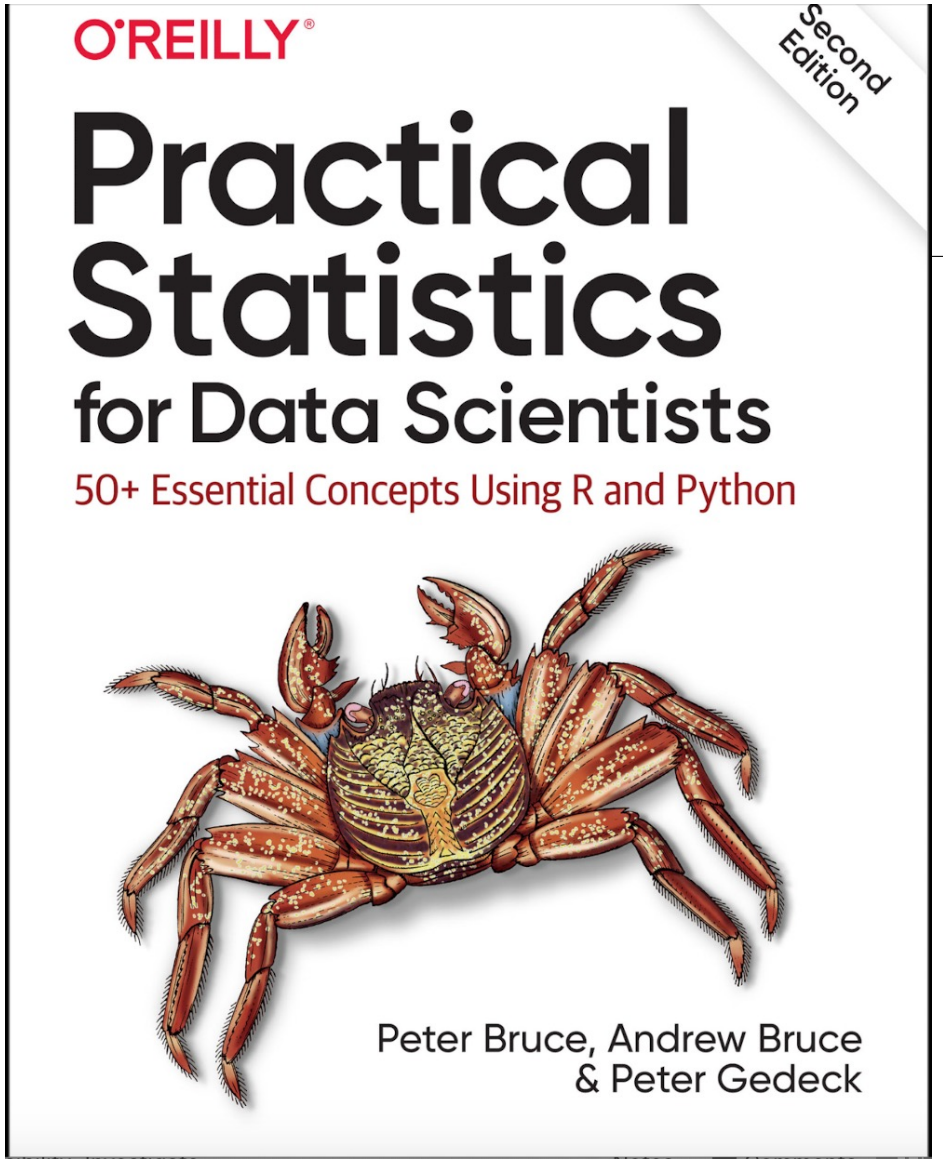
Lecture slides and examples committed to GitHub also up through lecture 7.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience

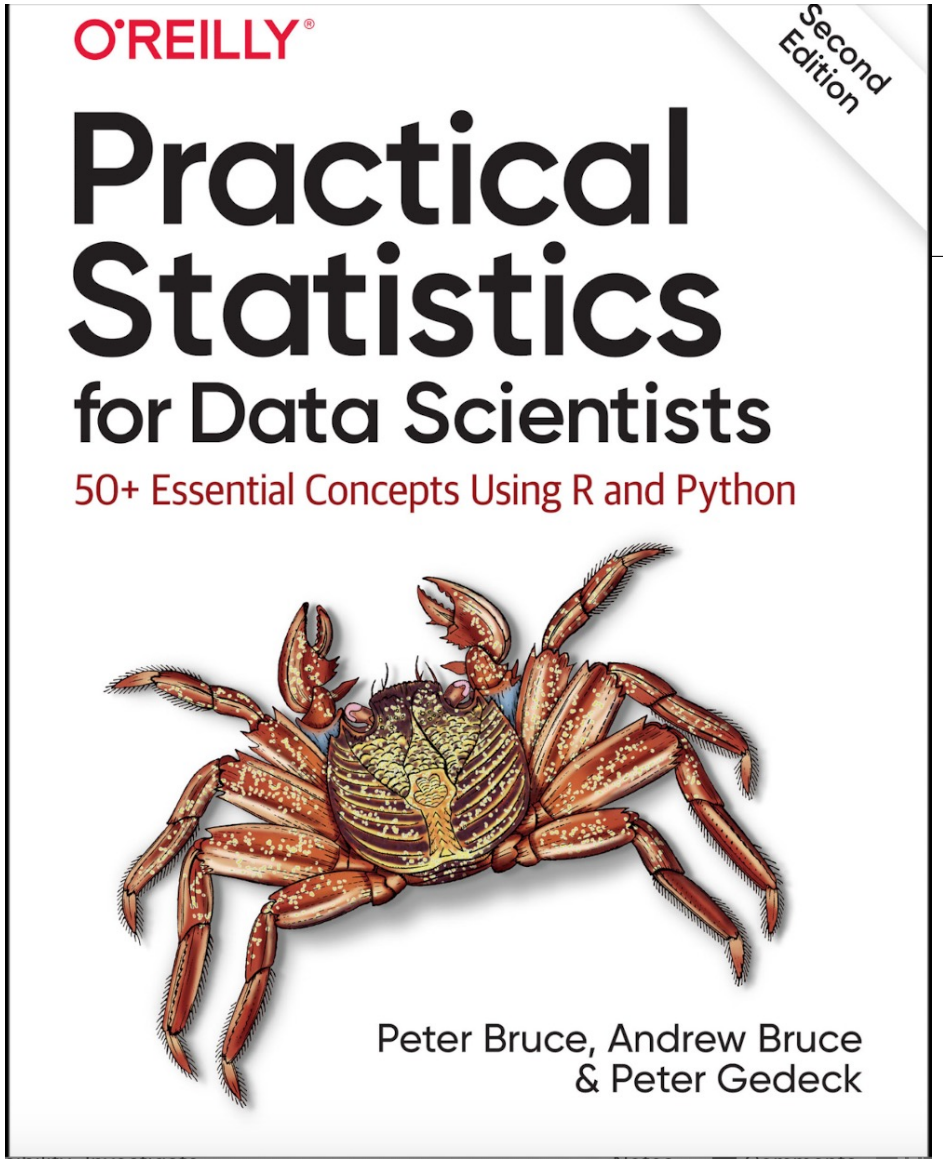
READ ABOUT

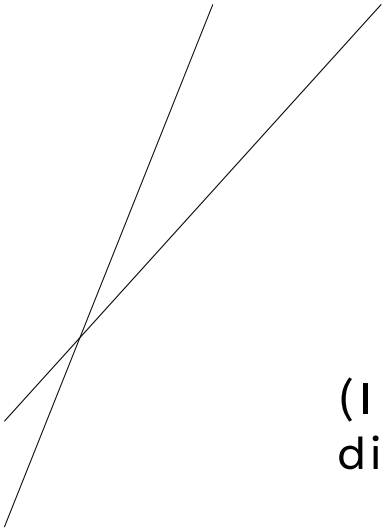
- Bias
  - Examples were already given in class, but book provides good example of selection bias.
- Random selection
  - ways to avoid bias
- Size vs. Quality: When Does Size Matter?



THINGS I WANT TO COVER TODAY

- Review Homework 2
- Chapter 2
  - How to evaluate Gaussian
    - without a computer
  - Z-scores





REVIEW HOMEWORK 2

(I will post the answers online, but for this discussion, let's use the whiteboard)

CUMULATIVE DISTRIBUTION FUNCTIONS

I want $P[X < x]$ where X is any random variable with given Probability Density Function $f(x)$.

To solve for this I would integrate the PDF.

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

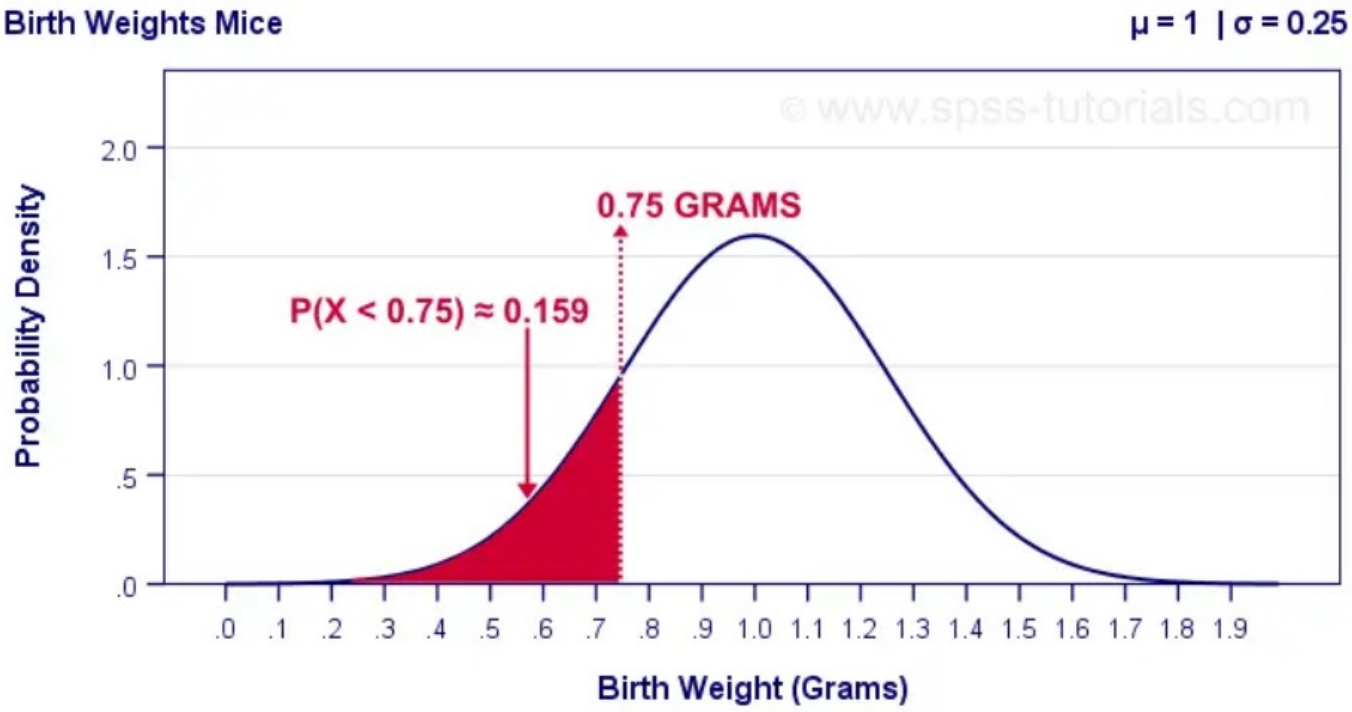
$F(x)$ is known as the *Cumulative Distribution Function* (CDF).

CUMULATIVE DISTRIBUTION FUNCTION EXAMPLE.

Find $P[X < x]$

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

Example finding percentiles of a distribution where $F(x) = 0.159$*, i.e.,* 15.9^{th} *percentile*



GAUSSIAN DISTRIBUTIONS

Let X = a Gaussian Random Variable. This is denoted

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

μ = mean
σ = std dev

Where N means "Normal" == "Gaussian"

Probability density function (PDF) $f(x)$ given by

Using the exp notation:

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

Using the e^{-} notation:

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

Both equations mean the same thing. e^x = exp(x)

GAUSSIAN DISTRIBUTIONS

Gaussian Probability Density Function.

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

To find the P that X < x we integrate the PDF

$$P[X \leq x] = F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

Problem: no known closed form solution for the CDF of a Gaussian random variable.

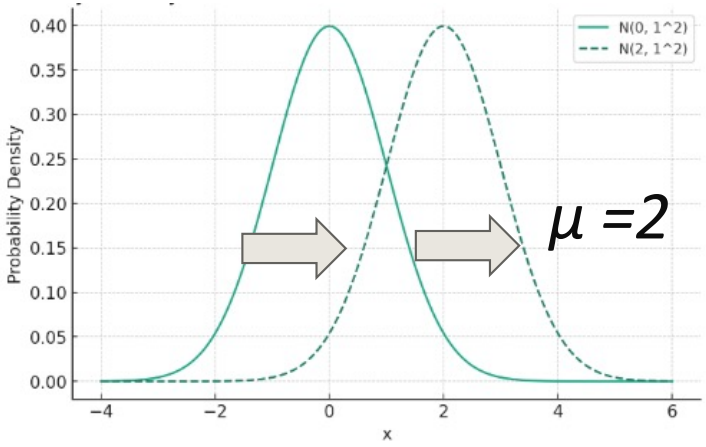
WE EVALUATE GAUSSIAN CDF NUMERICALLY

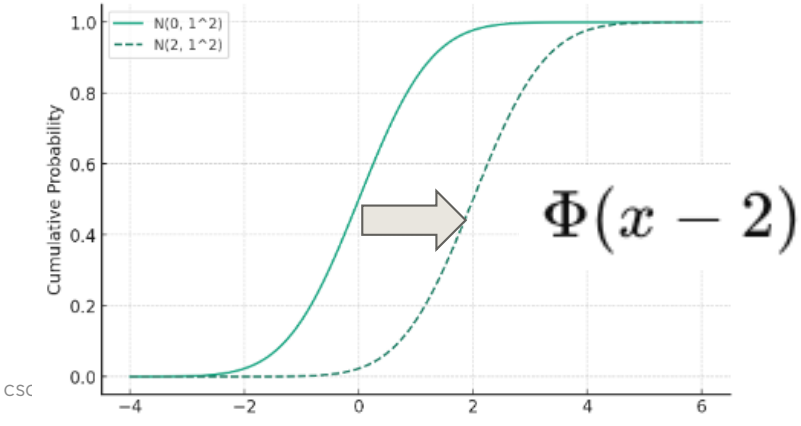
For Gaussian, we can find a table of the values for the CDF of a Gaussian with μ =0 and σ =1. This CDF is known as Φ (i.e., Phi).

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

We can transform to provide the CDF for a Gaussian of any mean, by shifting x by μ.

$$F(x) = \Phi(x - \mu)$$

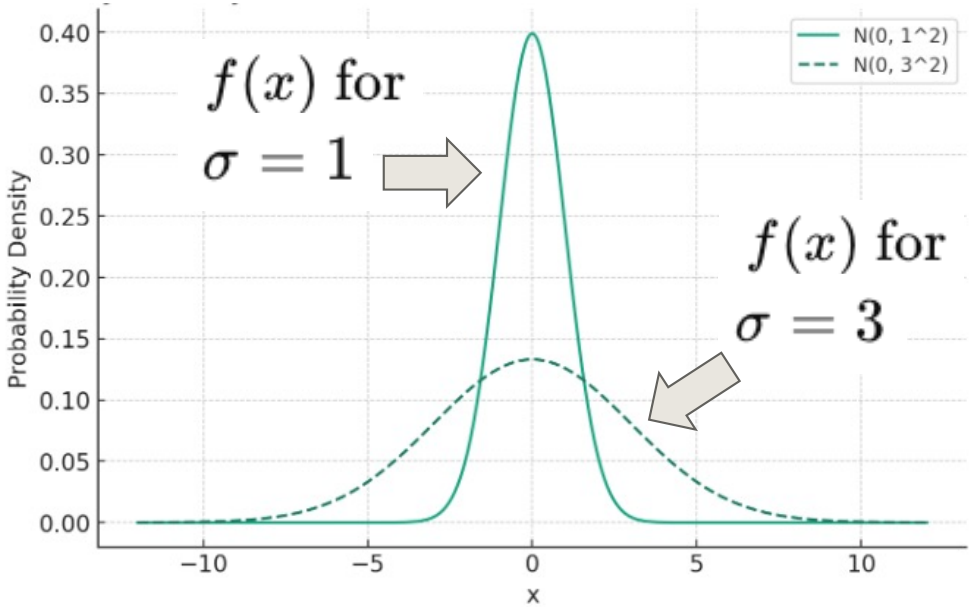


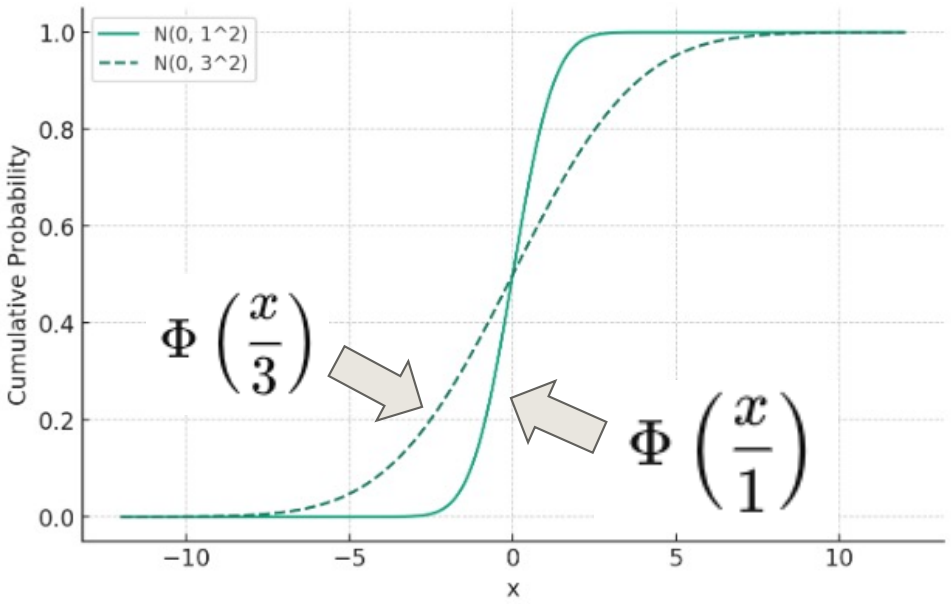


WE EVALUATE GAUSSIAN CDF NUMERICALLY

We can further transform $\Phi(x)$, Phi, to handle Gaussian distributions with any value for sigma as follows:

$$F\left(\frac{x}{\sigma}\right) = \Phi\left(\frac{x}{\sigma}\right)$$

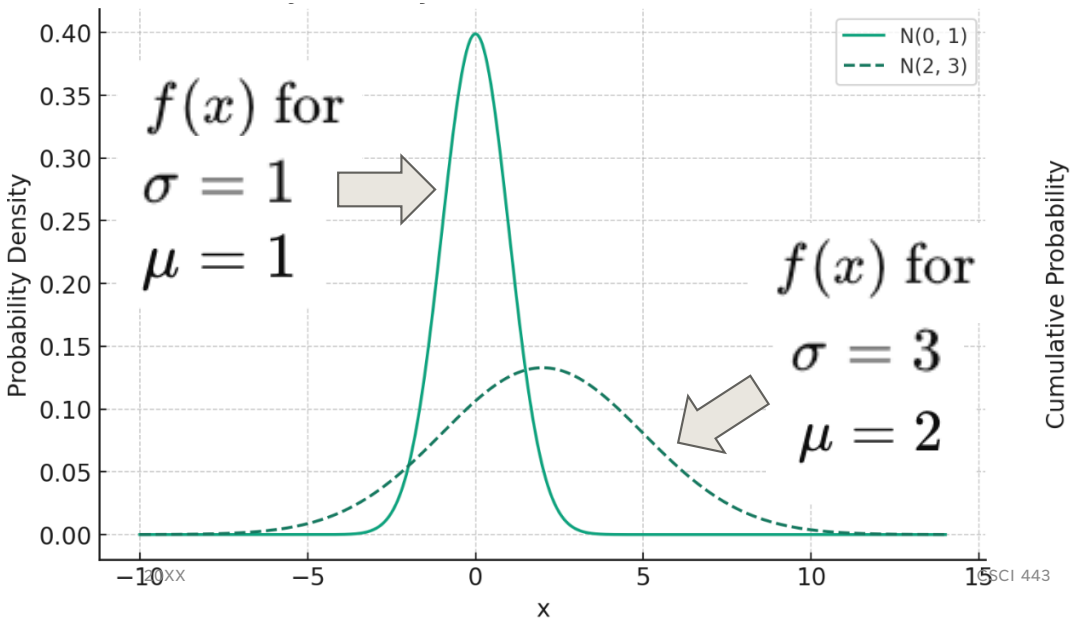


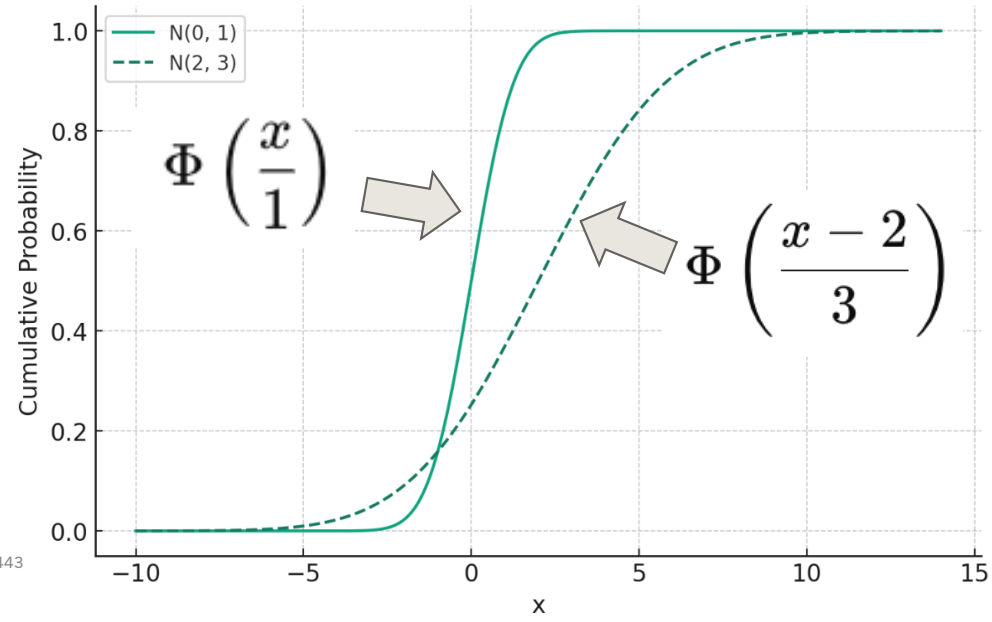


EVALUATING ARBITRARY GAUSSIAN FUNCTIONS

We can transform $\Phi(x)$ to handle arbitrary Gaussian distributions by combining the transforms for μ and σ as follows:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$





Z-SCORES

A Z-score tells you how far samples deviates from the mean in units of standard deviations. The Z-score for a sample

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

If a sample has a Z score of 1 then the sample is 1 standard deviation above the mean.

Z=-1 implies the sample it 1 standard deviation below the mean.

EXAMPLE

We can transform $\Phi(x)$ to handle arbitrary Gaussian distributions by combining the transforms for μ and σ as follows:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Let's consider with mean 45 and std dev 10.

John scored 70 on an exam. Enough people took the exam such that it looks like a Gaussian distribution as long as we aren't too close to 100 or 0. If there are 6475 people took the exam, approximately how many scored lower than John.

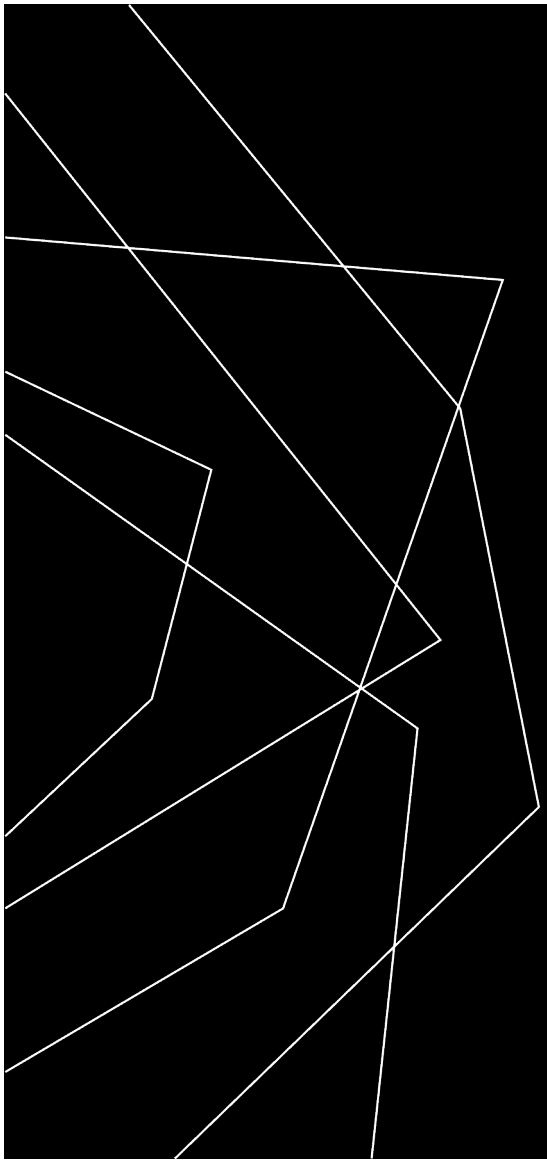
EXAMPLE

$$Z = 2.5$$

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

$\Phi(2.5)$ is approximately 0.9938.

He scored better than approximately 6434 people out of 6475.



THANK YOU

David Harrison

Harrison@cs.olemiss.edu