# Probability Theory

## Applications for Data Science
## Module 4 Continuous Random Variables

Anne Dougherty

March 7, 2021

# TABLE OF CONTENTS

# Random Variables

At the end of this module, students should be able to

- Define a continuous random variable and give examples of a probability density function and a cumulative distribution function.
- Identify and discuss the properties of a uniform, exponential, and **normal random variable**
- Calculate the expectation and variance of a continuous rv.

Part 1 – introduce normal r.v. & the standard normal. In part 2 we'll go over examples.

Normal (or Gaussian) distribution is probably the most important, and widely used, distribution in all of probability and statistics. – it has the typical bell shaped density fcn.

Many populations have distributions that can be fit very closely by an appropriate normal bell curve. which is one of the reasons for its usefulness and widespread applicability

Examples: height, weight, and other physical characteristics, scores on some tests, some error measurements, etc. can be modeled by a Gaussian distribution.

The normal rv has almost a 300 year history. It's been used by many statisticians + scientists.

## Normal (or Gaussian) random variable

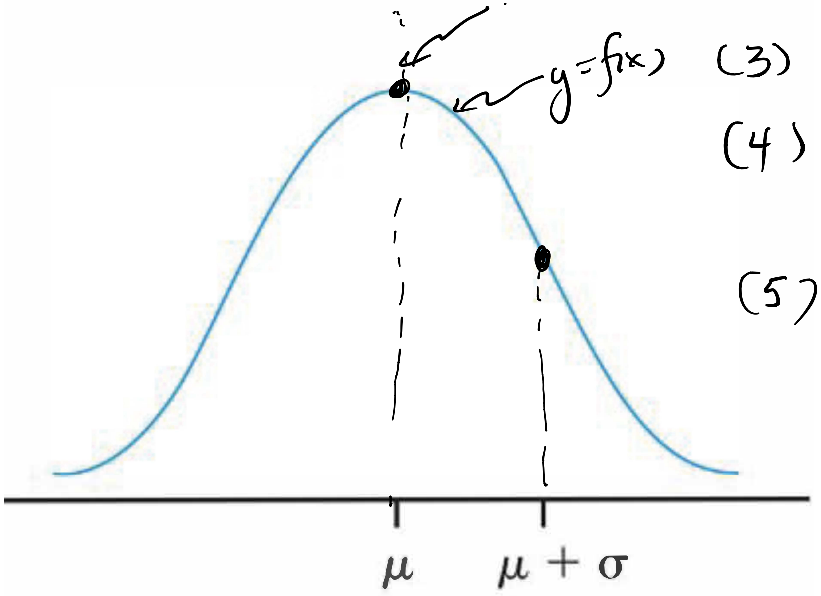
- First used by Abraham deMoivre in 1733, later by many others, including Carl Friedrich Gauss.
- Gauss used it so extensively in his astronomical calculations, it came to be called the Gaussian distribution. Karl Pearson
- In 1893, Karl E. Pearson wrote "Many years ago I called the Laplace-Gaussian curve the **normal curve**, which name, while it avoids the international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another abnormal.".

Usually in math + science, the first person to discover something has it named after himself or herself. Pearson was trying to avoid this.

Definition: A continuous random variable $X$ has the normal distribution with parameters $\mu$ and $\sigma^2$ if its density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \text{ for } -\infty < x < \infty$$

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



Important properties

(1) $f(x)$ is symmetric around the line $x = \mu$

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

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

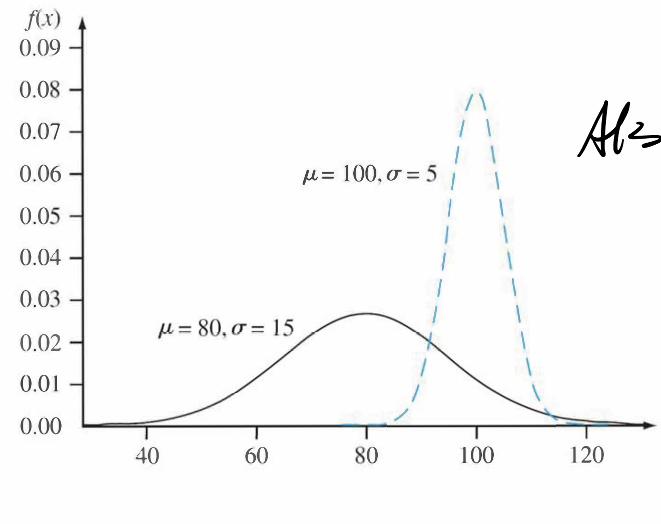
(4) $V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$
$= \sigma^2$

(5) $\sigma$ = standard deviation and $\mu + \sigma$ and $\mu - \sigma$ are inflection points for $y = f(x)$

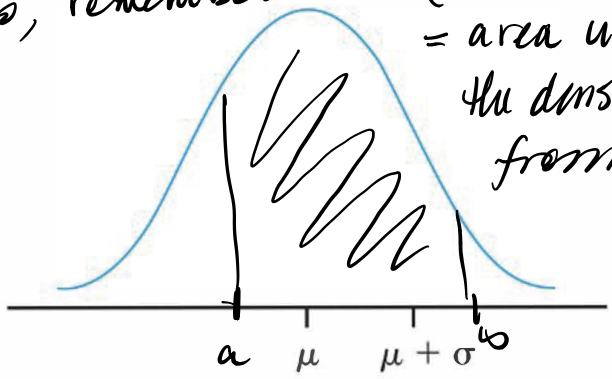
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \text{ for } -\infty < x < \infty$$

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

Note: for smaller $\sigma$ the density fcn is more peaked. For larger $\sigma$, density fcn is more spread out.



Also, remember: $P(a \le X \le b)$ = area under the density fcn from a to b.



For both theoretical + practical reasons, it's usually easier to work with 'standard normal'

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

A rv with the standard normal distribution is customarily denoted by $Z \sim N(0,1)$ and its pdf is given by

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \text{ for } -\infty < x < \infty$$

We use special notation to denote the cdf of the standard normal curve:

$$F(z) = \Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

- The standard normal density function is symmetric about the $y$ axis.

- The standard normal distribution rarely occurs naturally.

- Instead, it is a reference distribution from which information about other normal distributions can be obtained via a simple formula.

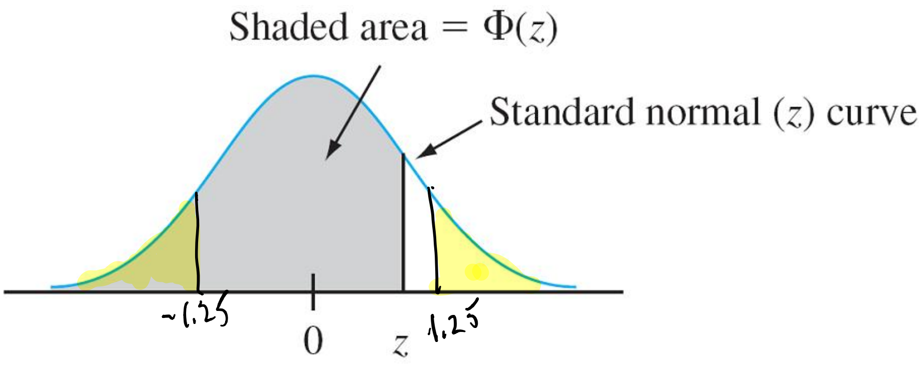
- The cdf of the standard normal, $\Phi$, can be found in tables and it can also be computed with a single command in R.

→ later in this course & in the next one

- As we'll see, sums of standard normal random variables play a large role in statistical analysis.

Example calculations:

- $P(Z \geq 1.25) = 1 - P(Z \leq 1.25) = 1 - \Phi(1.25) = .1056$
  $= \int_{1.25}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$
- Why does $\underbrace{P(Z \leq -1.25)}_{\Phi(-1.25)} = \underbrace{P(Z \geq 1.25)}_{1 - \Phi(1.25)}$? symmetry ~~$\Phi(-1.25) = 1 - \Phi(1$~~
- $P(-.38 \leq Z \leq 1.25) = P(Z \leq 1.25) - P(Z \leq -.38)$
  $= \Phi(1.25) - \Phi(-.38)$



- $\overbrace{P(-1 \leq Z \leq 1)}^{\text{Prob that } Z \text{ is within 1 std deviation of the mean}} = P(Z \leq 1) - P(Z \leq -1) = \Phi(1) - \Phi(-1)$
  $= .6826$

- $P(-2 \leq Z \leq 2) = P(Z \leq 2) - P(Z \leq -2)$
  $= \Phi(2) - \Phi(-2) = .9544$

Critical Values: In statistical inference, we need the $z$ values that give certain tail areas under the standard normal curve. For example, find $\overset{z}{\alpha}$ so that $\Phi(z_\alpha) = P(Z \leq z_\alpha) = .95$

Find $z_\alpha = 1.645$

$P(Z \leq 1.645) = .95$

and $P(-1.645 \leq Z \leq 1.645) = .90$

In the next video we'll see the relationship between a normal r.v. with mean $\mu$ & variance $\sigma^2$ and a standard normal & we'll work out some examples