The Gaussian (normal) Random Variable

Probability Theory: Foundation for Data Science

with **Anne Dougherty**

Data Science
UNIVERSITY OF COLORADO **BOULDER**

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.

Normal (or Gaussian) distribution is probably the most important, and widely used, distribution in all of probability and statistics.

Many populations have distributions that can be fit very closely by an appropriate normal bell curve.

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

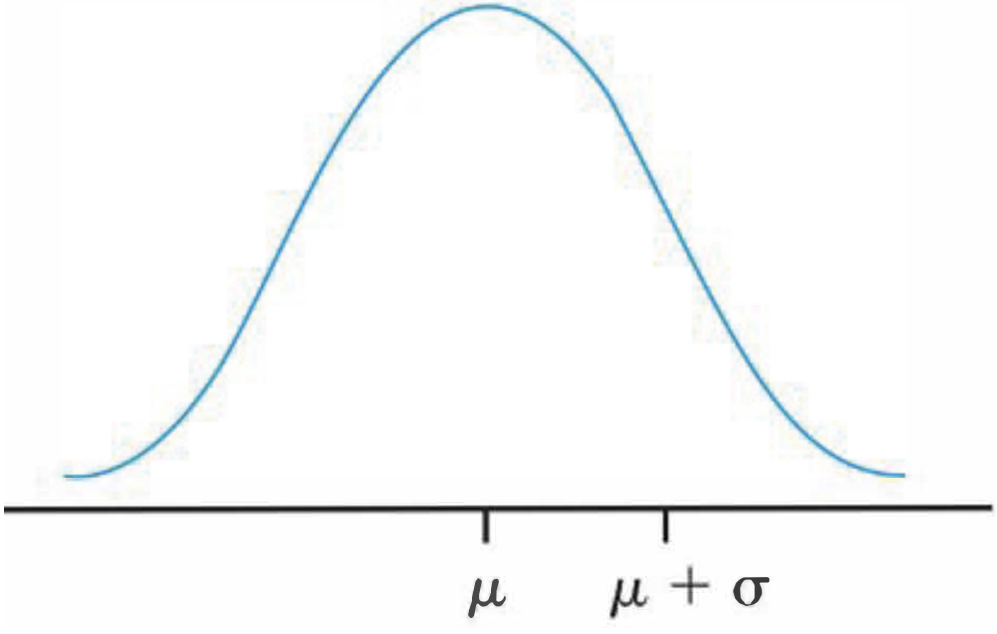
Normal (or Gaussian) random variable

- 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.
- 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."

Definition: A continuous random variable X has the normal distribution with parameters μ and σ^2 if its density is given by

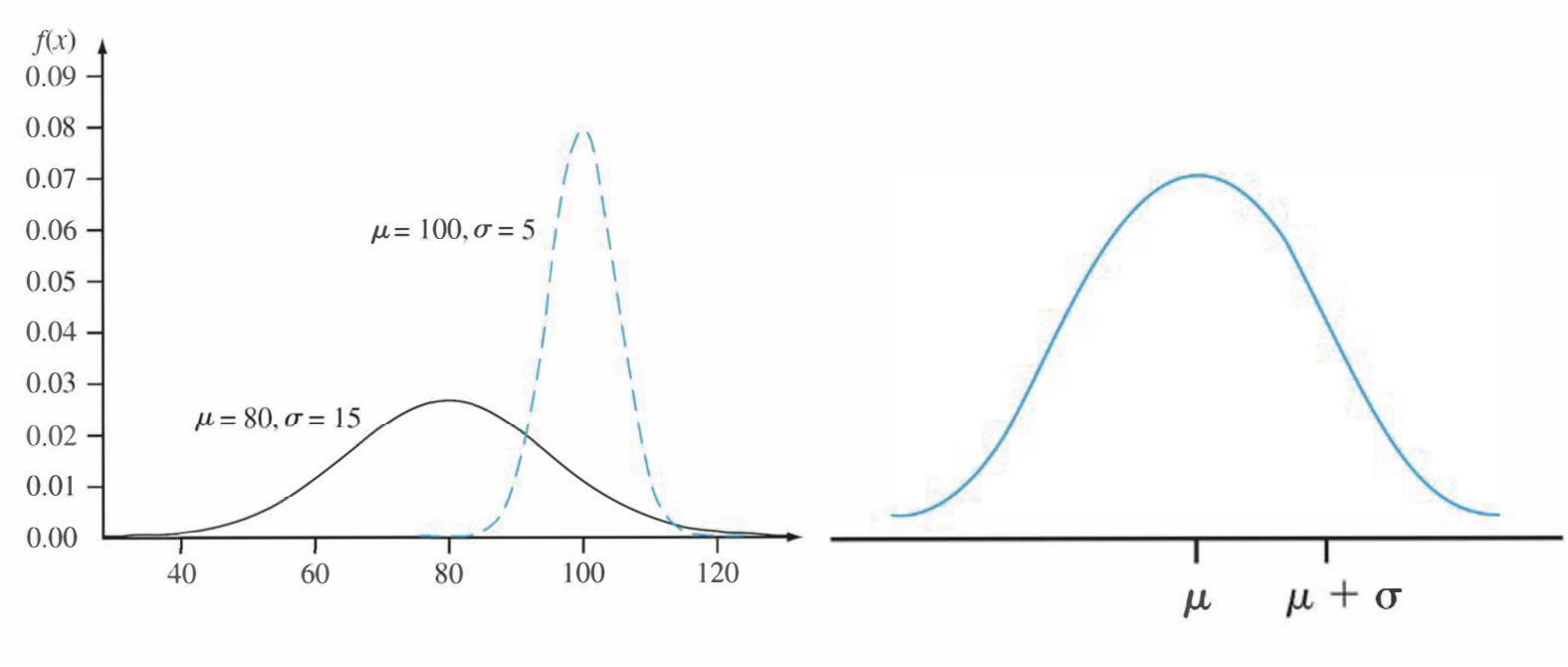
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \text{ for } -\infty < x < \infty$$

Notation:



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

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



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

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

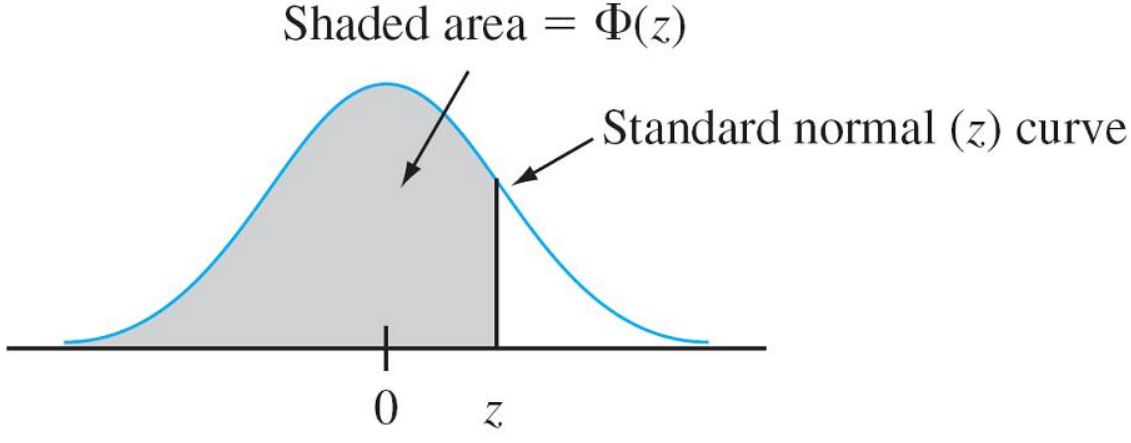
We use special notation to denote the cdf of the standard normal curve:

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

- 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, Φ, can be found in tables and it can also be computed with a single command in R.
- As we'll see, sums of standard normal random variables play a large role in statistical analysis.

Example calculations:

- $P(Z \geq 1.25) =$
- Why does $P(Z \leq -1.25) = P(Z \geq 1.25)$?
- $P(-.38 \leq Z \leq 1.25) =$



- $P(-1 \leq Z \leq 1)$
- $P(-2 \leq Z \leq 2)$

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