

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2009, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL of PUBLIC HEALTH

## Describing Data: Part II

---

John McGready  
Johns Hopkins University

# Lecture Topics

- The normal distribution
- Means, variability, and the normal distribution
- Calculating normal (z) scores
- Means, variability and z-scores for non-normal distributions



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

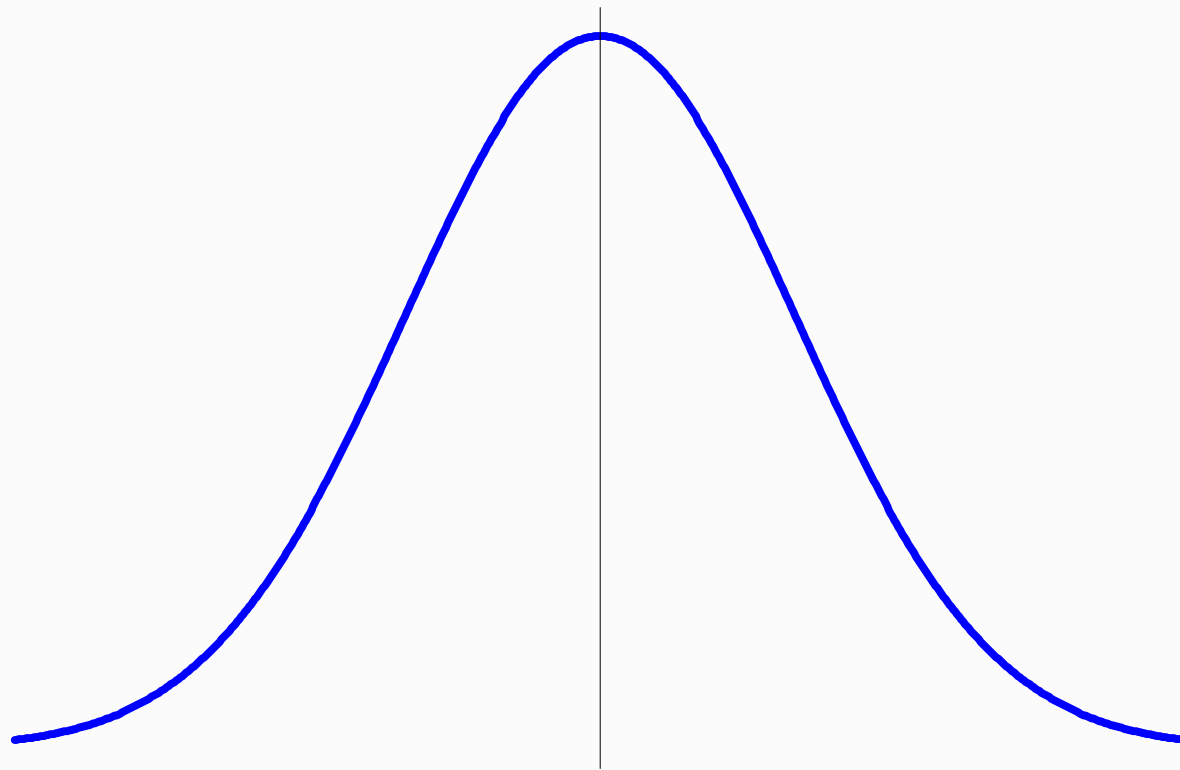
## Section A

---

### The Normal Distribution

# The Normal Distribution

- The **normal distribution** is a theoretical probability distribution that is perfectly symmetric about its mean (and median and mode), and had a “bell” like shape



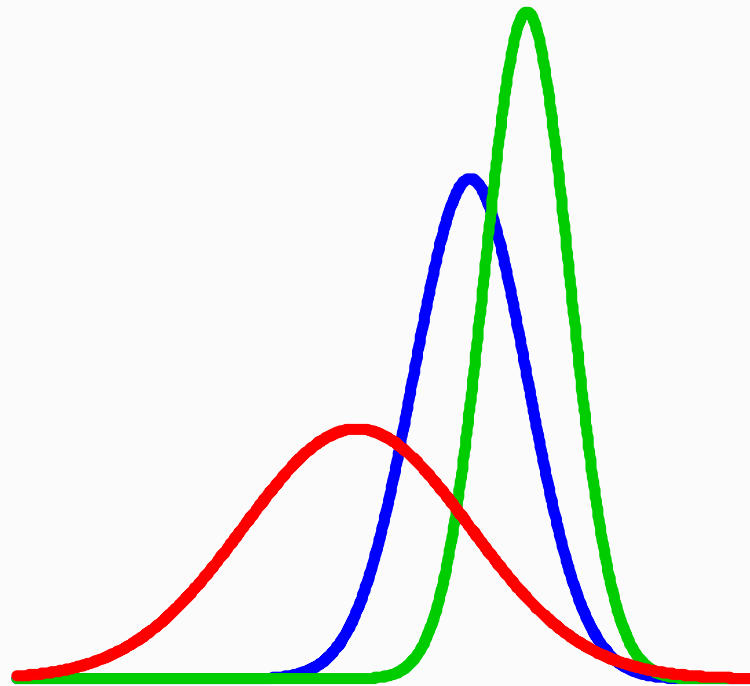
# The Normal Distribution

- The normal distribution is also called the “Gaussian distribution” in honor of its inventor Carl Friedrich Gauss



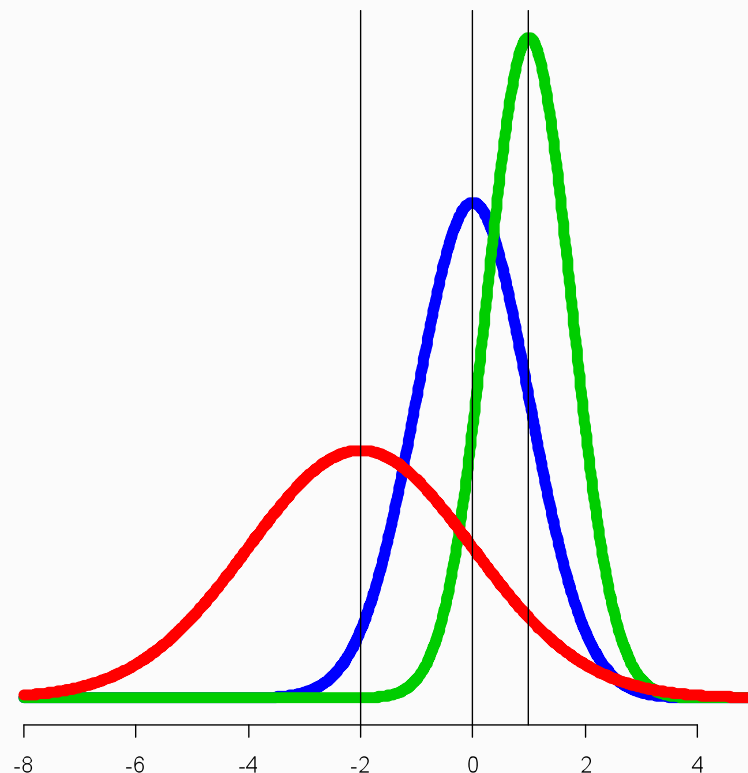
# The Normal Distribution

- Normal distributions are uniquely defined by two quantities: a mean ( $\mu$ ), and standard deviation ( $\sigma$ )
- There are literally an infinite number of possible normal curves, for every possible combination of ( $\mu$ ) and ( $\sigma$ )



# The Normal Distribution

- Normal distributions are uniquely defined by two quantities: a mean ( $\mu$ ), and standard deviation ( $\sigma$ )
- There are literally an infinite number of possible normal curves, for every possible combination of ( $\mu$ ) and ( $\sigma$ )





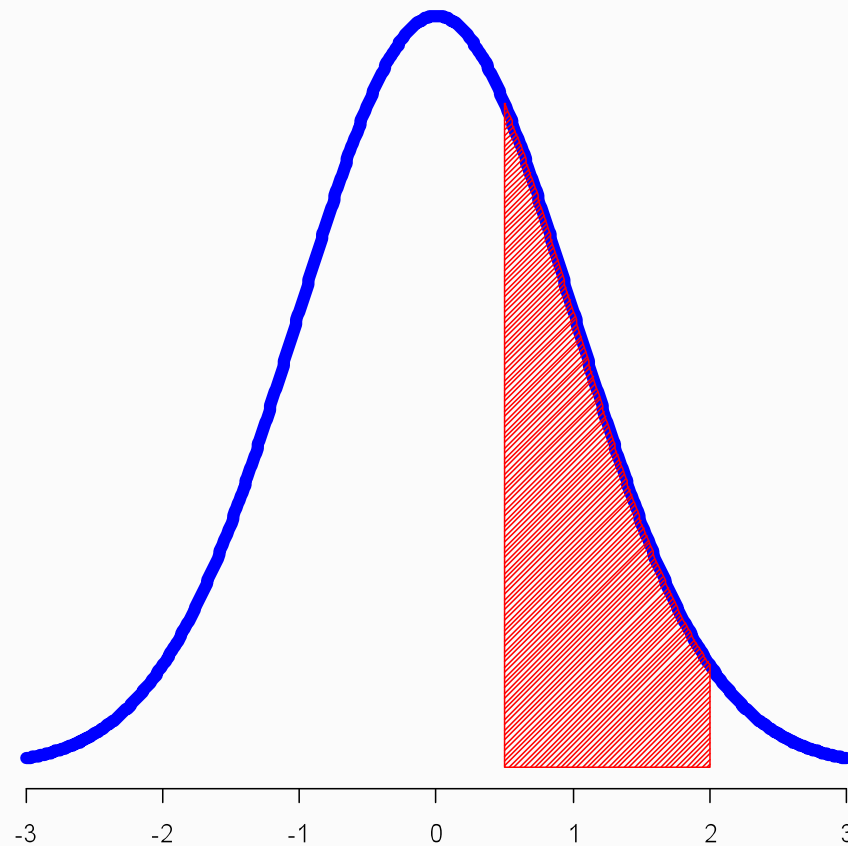
# The Normal Distribution

- Normal distributions are uniquely defined by two quantities: a mean ( $\mu$ ), and standard deviation ( $\sigma$ )
- There are literally an infinite number of possible normal curves, for every possible combination of ( $\mu$ ) and ( $\sigma$ )
- This function defines the normal curve for any given ( $\mu$ ) and ( $\sigma$ )

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

# Normal Distribution

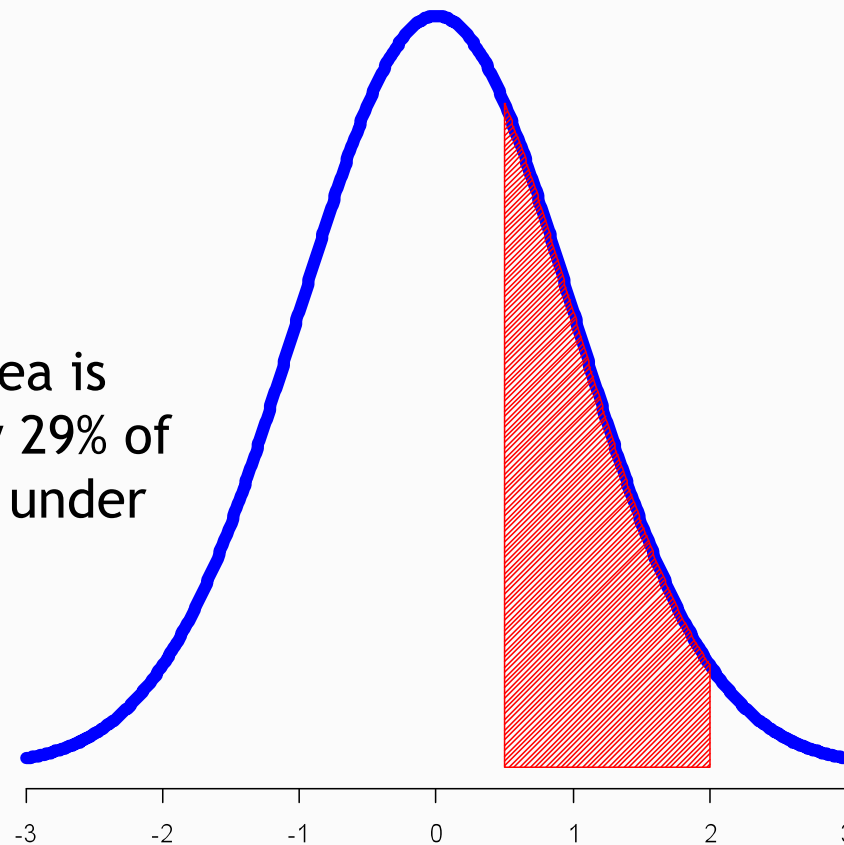
- Areas under a normal curve represent the proportion of total values described by the curve that fall in that range



# Normal Distribution

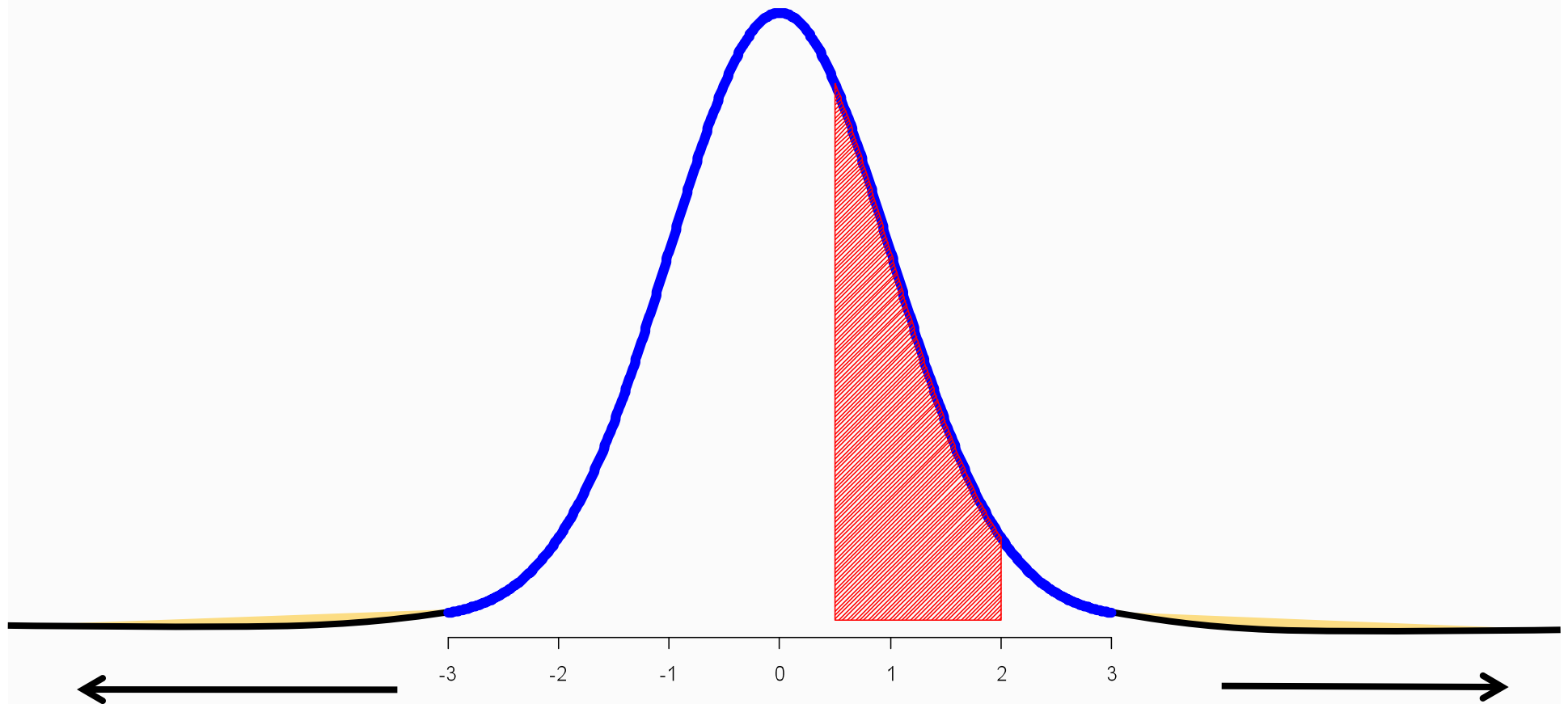
- This shaded area represents the proportion of values (observations) between 0 and 1 following a normal distribution with  $\mu = 0$  and  $\sigma = 1$

- The shaded area is approximately 29% of the total area under the curve



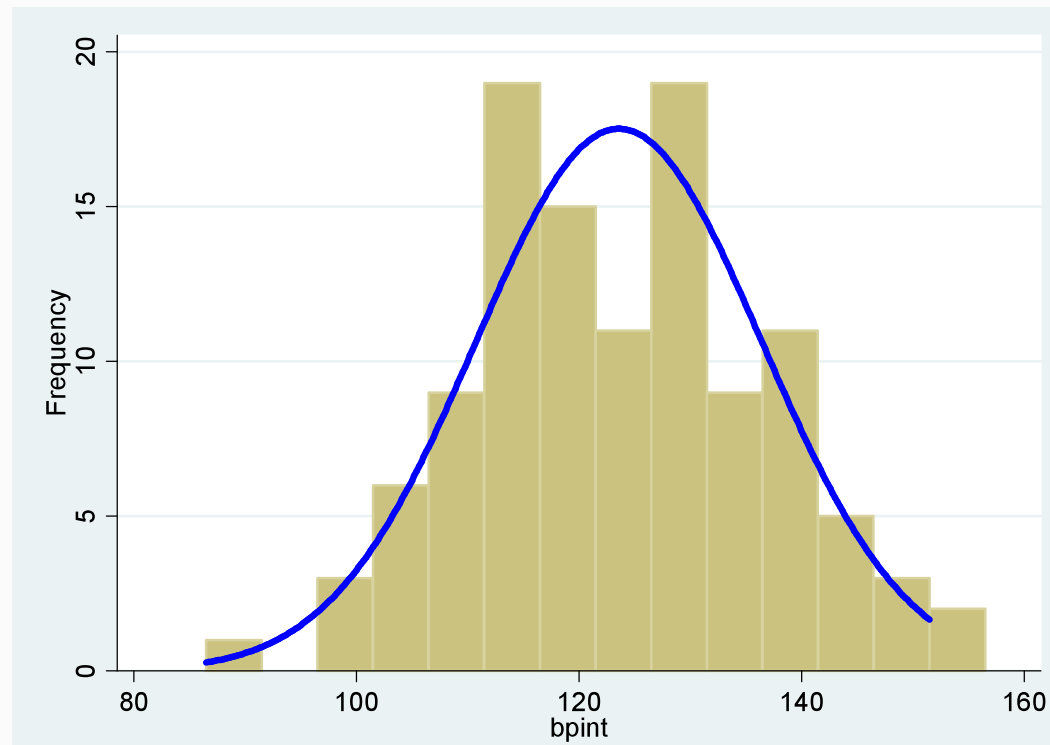
# Normal Distribution

- The normal distribution is a theoretical distribution: no real data will truly be normally distributed (at the sample or population level)
  - For example: the tails of the normal curve are “infinite”



# Normal Distribution

- BUT: some data approximates a normal curve pretty well
- Here is a histogram of the BP of the 113 men with a normal curve superimposed (normal curve has same mean and SD as sample of 113 men)
  - Mean 123.6 mmHg, SD 12.9 mmHg



# Normal Distribution

- Other data, does not approximate a normal distribution
- Here is a histogram of the hospital length of stay of the 500 patients with a normal curve superimposed (normal curve has same mean and SD as sample of 500 patients)
  - Mean 5.1 days, SD 6.4 days

