

定量分析：数据思维与 商业统计

陈文波

cwb@whu.edu.cn

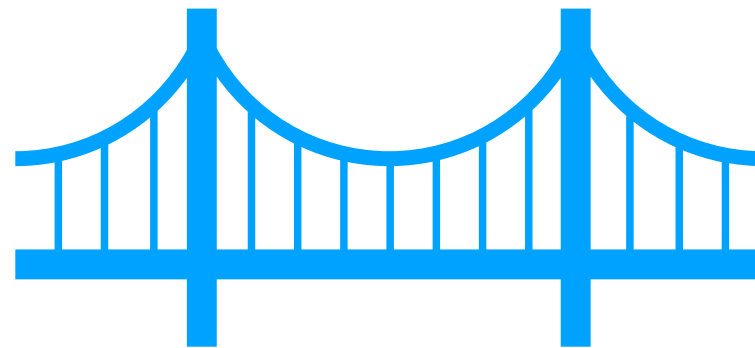
2021年10月

课程梗概

描述性分析与EDA

- 数值方法：三类指标
- 图形方法：五种图形

R



- 抽样分布、CLT
- 正态分布
- T分布
- 卡方分布
- F分布

推断性分析

- 区间估计
- 假设检验
- 应用：
 - 分类 vs. 分类
 - 分类 vs. 数值
 - 数值 vs. 数值

课程梗概

- Session 1 : 内容简介、基本概念、描述性统计1
- Session 2: 描述性统计2
- Session 3: 抽样、中心极限定理、正态分布
- Session 4: 其他三个分布, 区间估计
- Session 5: 假设检验, 两个总体均值和比例的推断
- Session 6: 总体方差和方差比的推断、分类数据 Vs. 数值型数据: 卡方检验与列联表分析
- Session 7: 分类数据 Vs. 数值型数据: 方差分析; 数值型数据: 回归分析
- Session 8: 课程总结

S3-抽样、中心极限定理与正态分布

- Review: 分布、如何描述分布
- 抽样方法
- 抽样分布
- 中心极限定理
- 正态分布

分布以及如何描述分布

- 概率分布 (probability distribution)的简称, 一个随机变量在一定范围内取值与其对应的可能性 (概率) 的关系的图、表或数学函数表达式。
- 图
- 表
- 数学表达式



抽样

- 抽样
- 抽样分布
- 四种随机抽样方式

中心极限定理

Central Limit Theorem

- 有关样本均值抽样分布

CENTRAL LIMIT THEOREM

In selecting random samples of size n from a population, the sampling distribution of the sample mean \bar{x} can be approximated by a *normal distribution* as the sample size becomes large.

正态分布

- The normal probability distribution is the most important distribution for describing a continuous random variable.
- It is widely used in statistical inference.
- It has been used in a wide variety of applications including:
 - Heights of people
 - Test scores
 - Amounts of rainfall
 - Scientific measurements
- Abraham de Moivre, a French mathematician, published *The Doctrine of Chances* in 1733.
- He derived the normal distribution.

正态分布

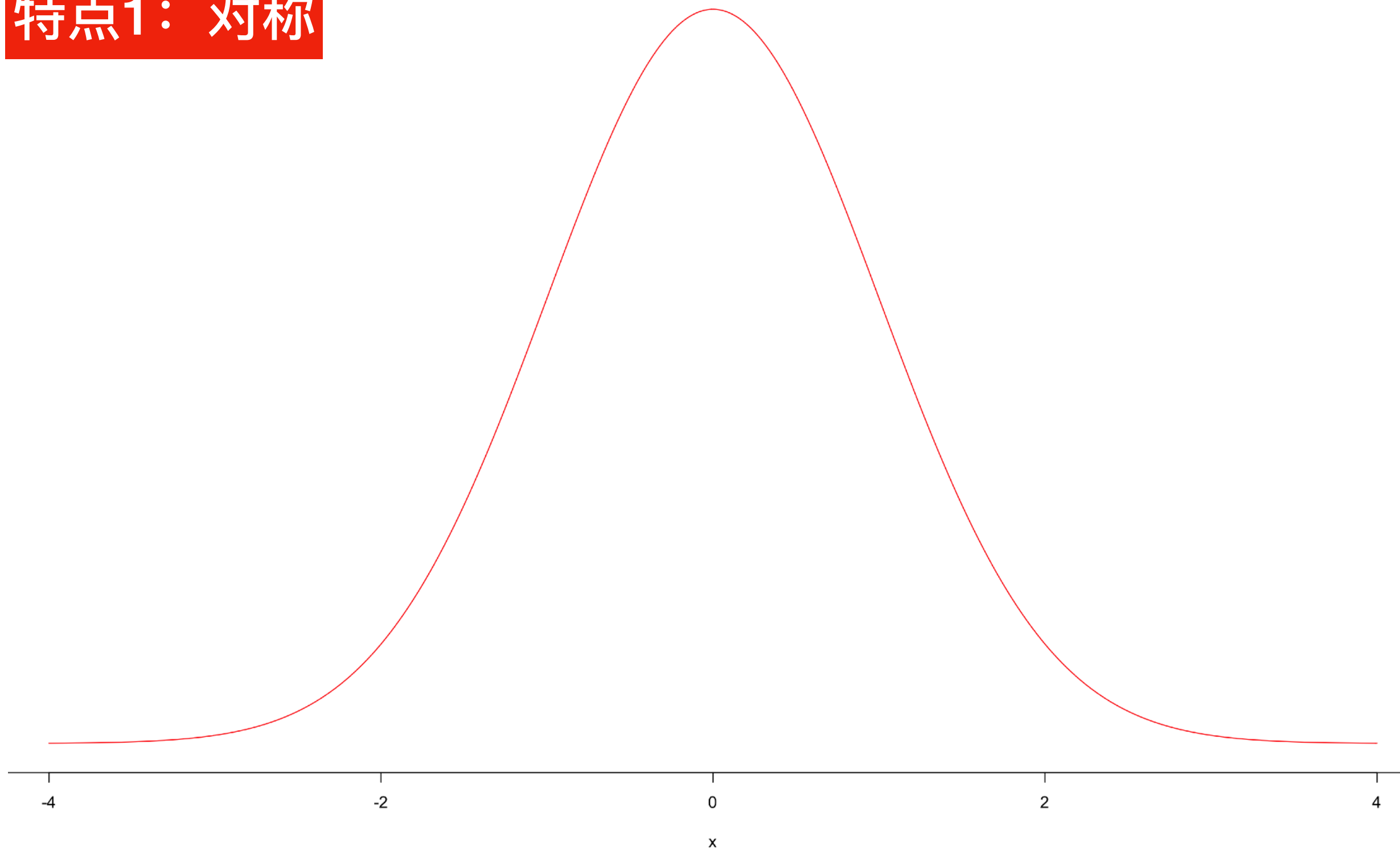
- Normal Probability Density Function

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

where: μ = mean
 σ = standard deviation
 π = 3.14159
 e = 2.71828

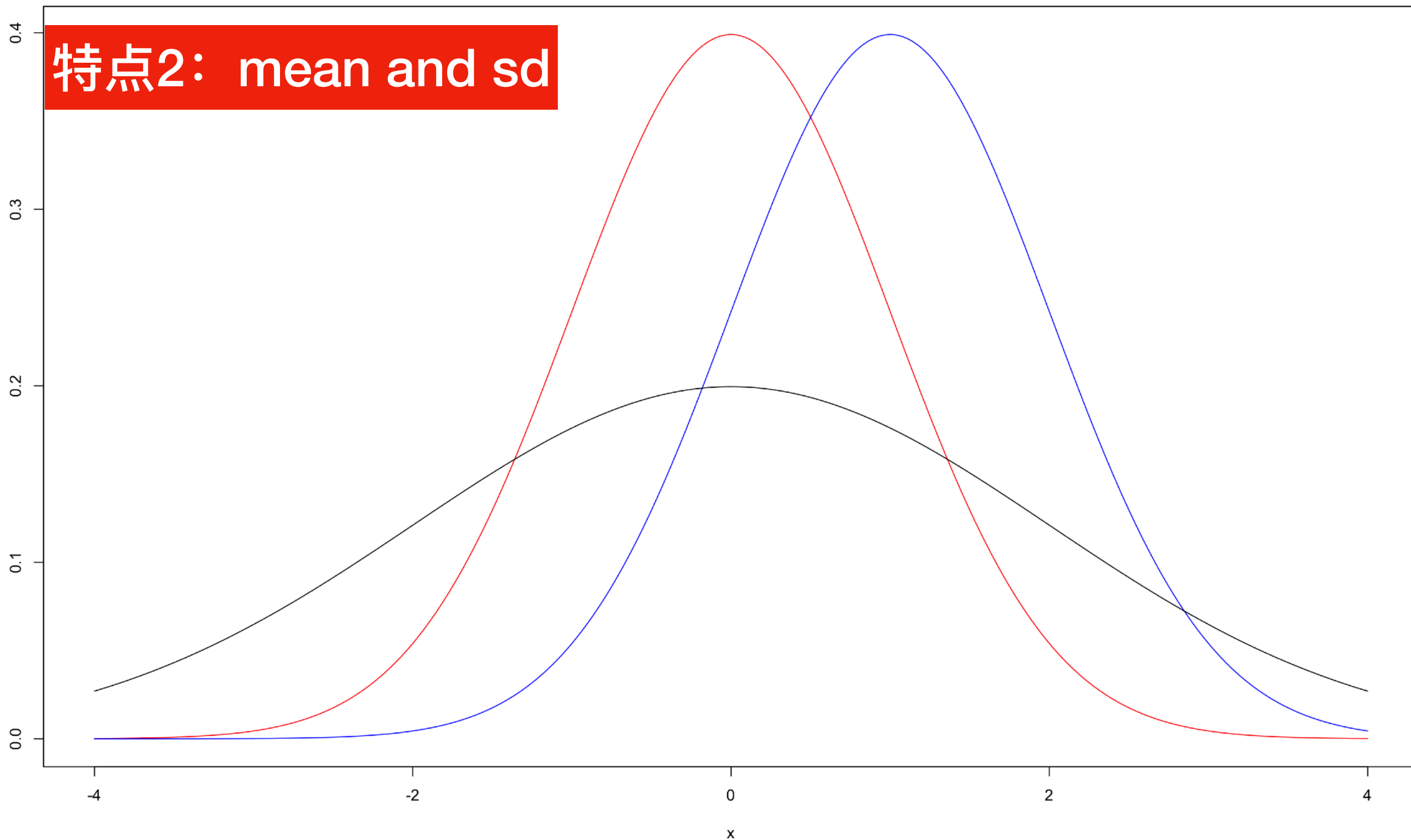
正态分布

特点1：对称



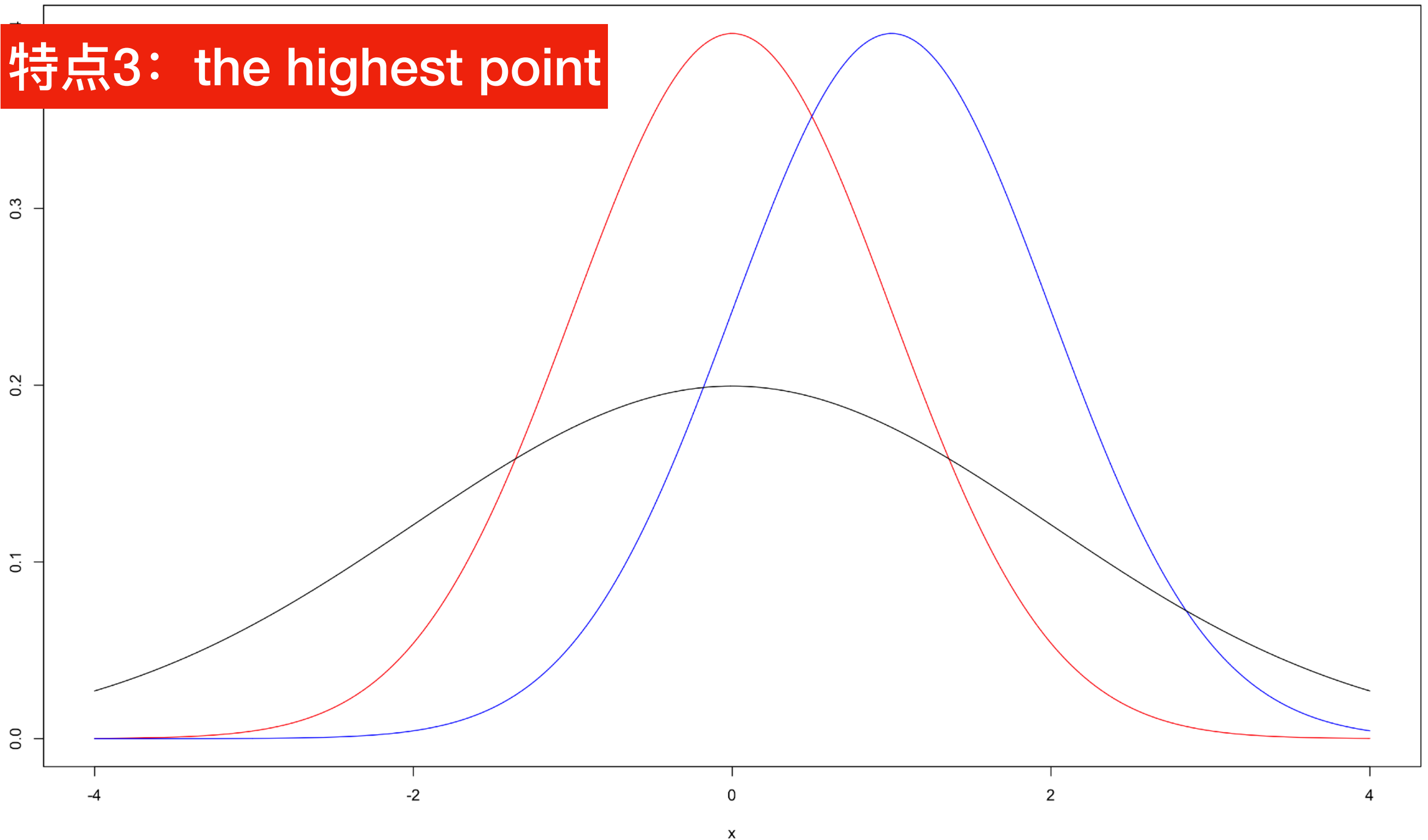
正态分布

特点2: mean and sd



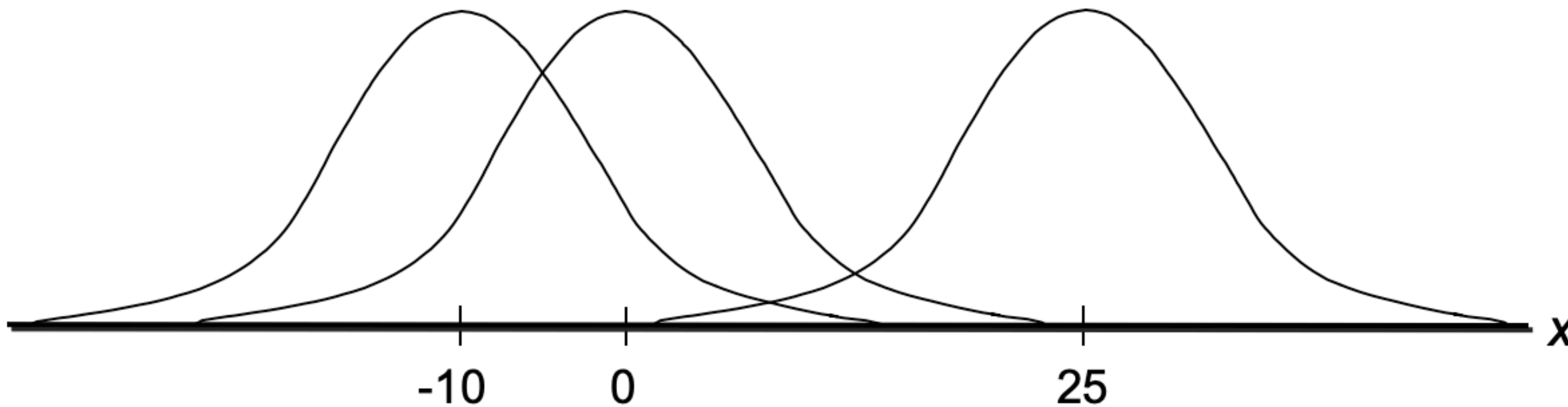
正态分布

特点3: the highest point



正态分布

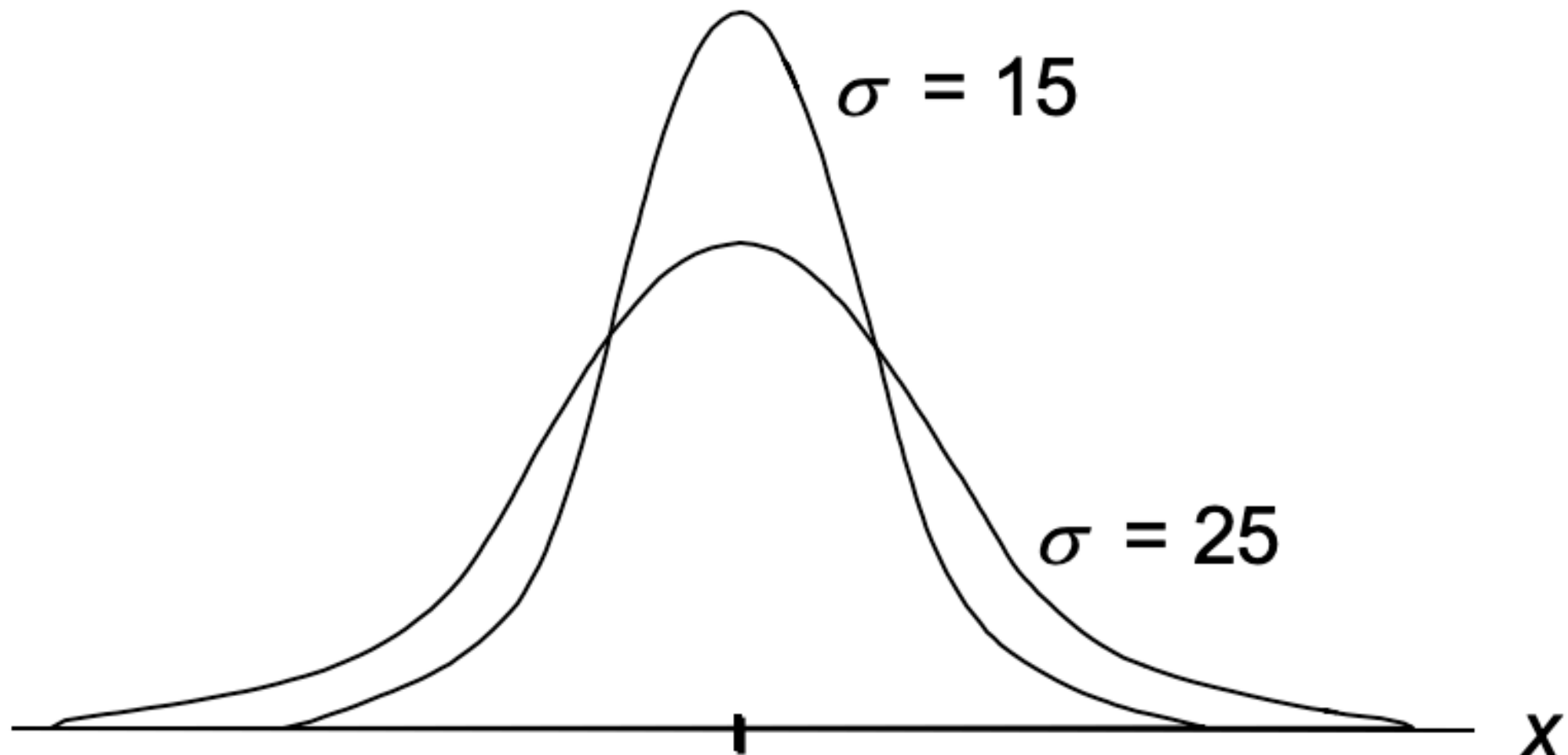
The mean can be any numerical value: negative, zero, or positive.



正态分布

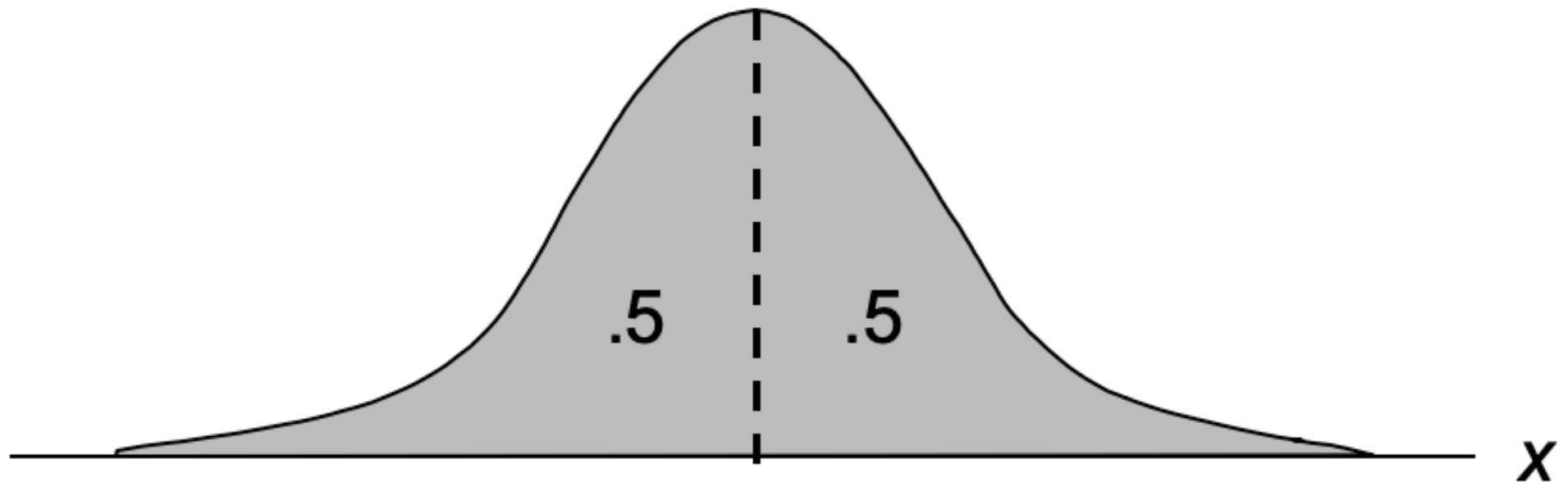
Characteristics

The standard deviation determines the width of the curve: larger values result in wider, flatter curves.



正态分布

Probabilities for the normal random variable are given by areas under the curve. The total area under the curve is 1 (.5 to the left of the mean and .5 to the right).



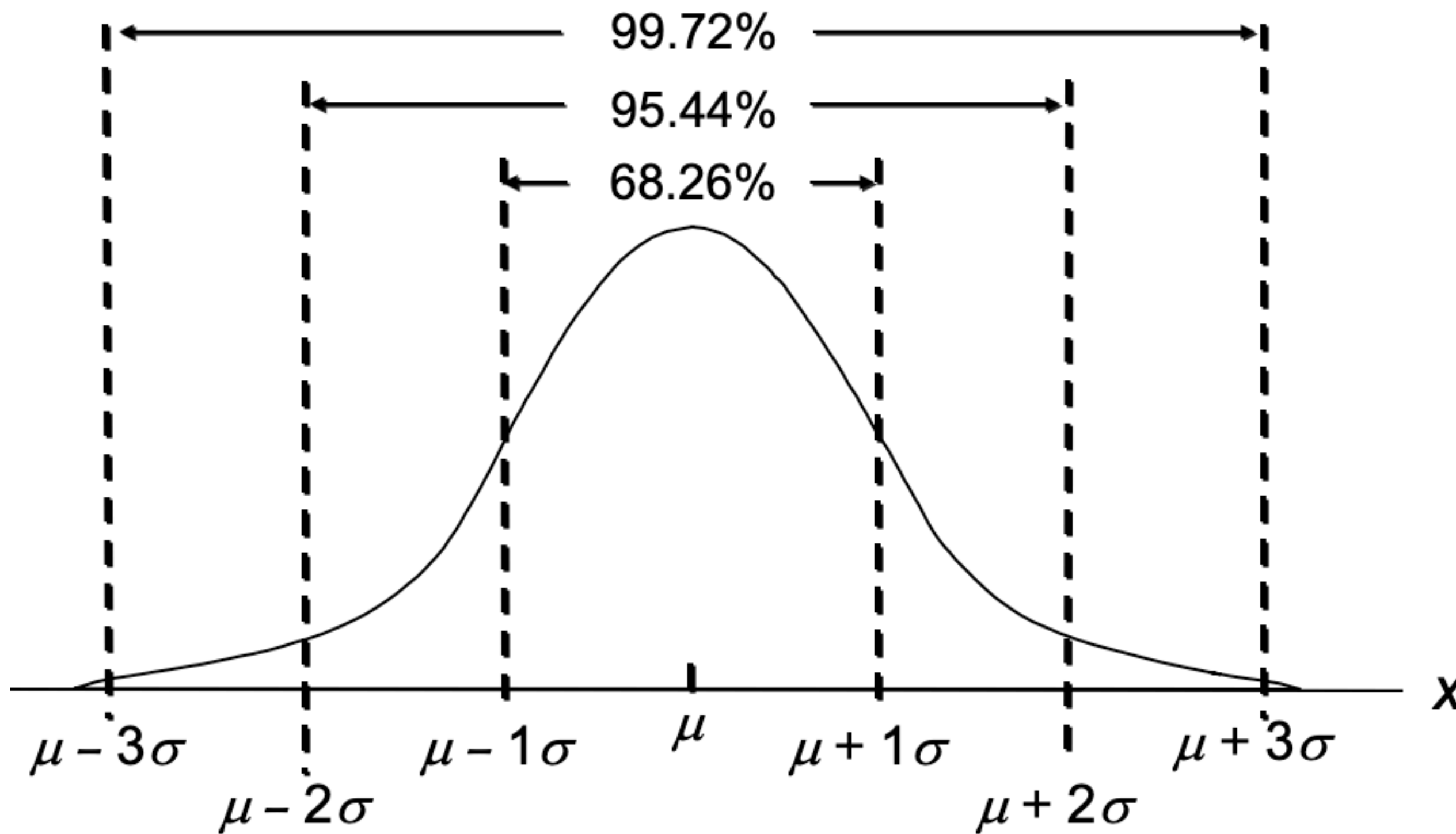
Empirical Rule

68.26% of values of a normal random variable are within ± 1 standard deviation of its mean.

95.44% of values of a normal random variable are within ± 2 standard deviations of its mean.

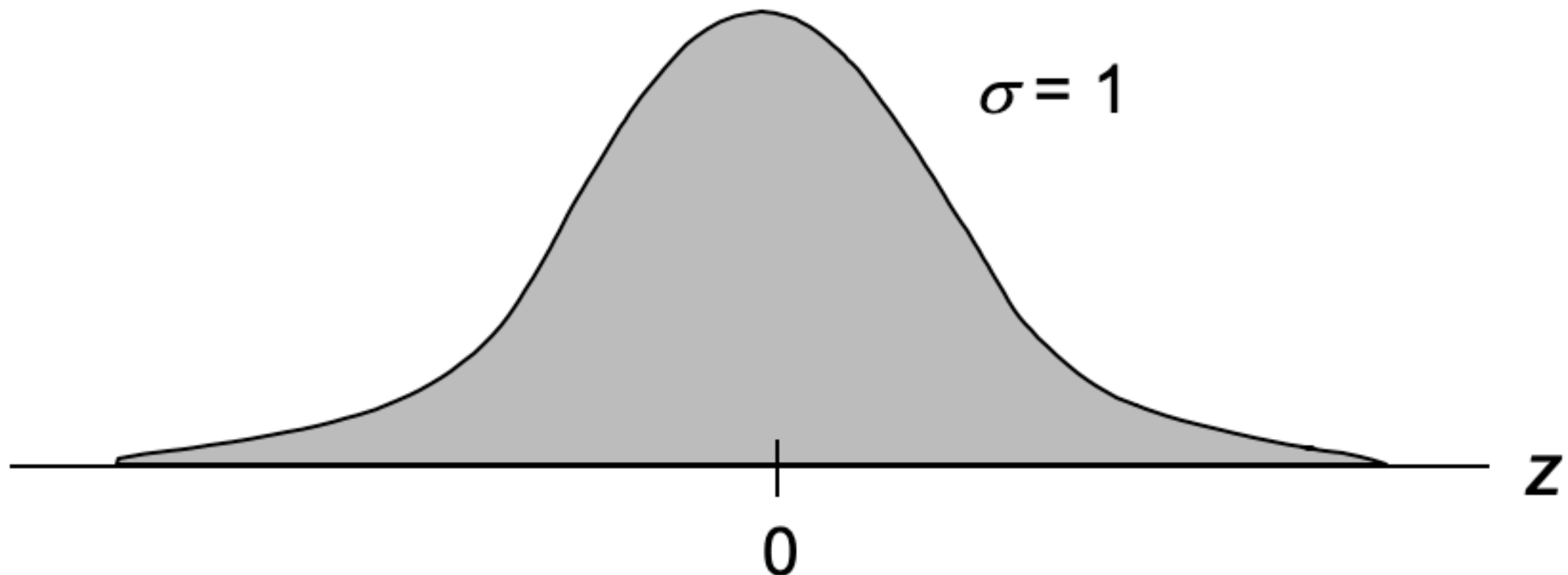
99.72% of values of a normal random variable are within ± 3 standard deviations of its mean.

正态分布



正态分布

- 标准正态分布：均值为0，标准差为1的正态分布
- 常用 Z 来表示标准正态分布变量



正态分布

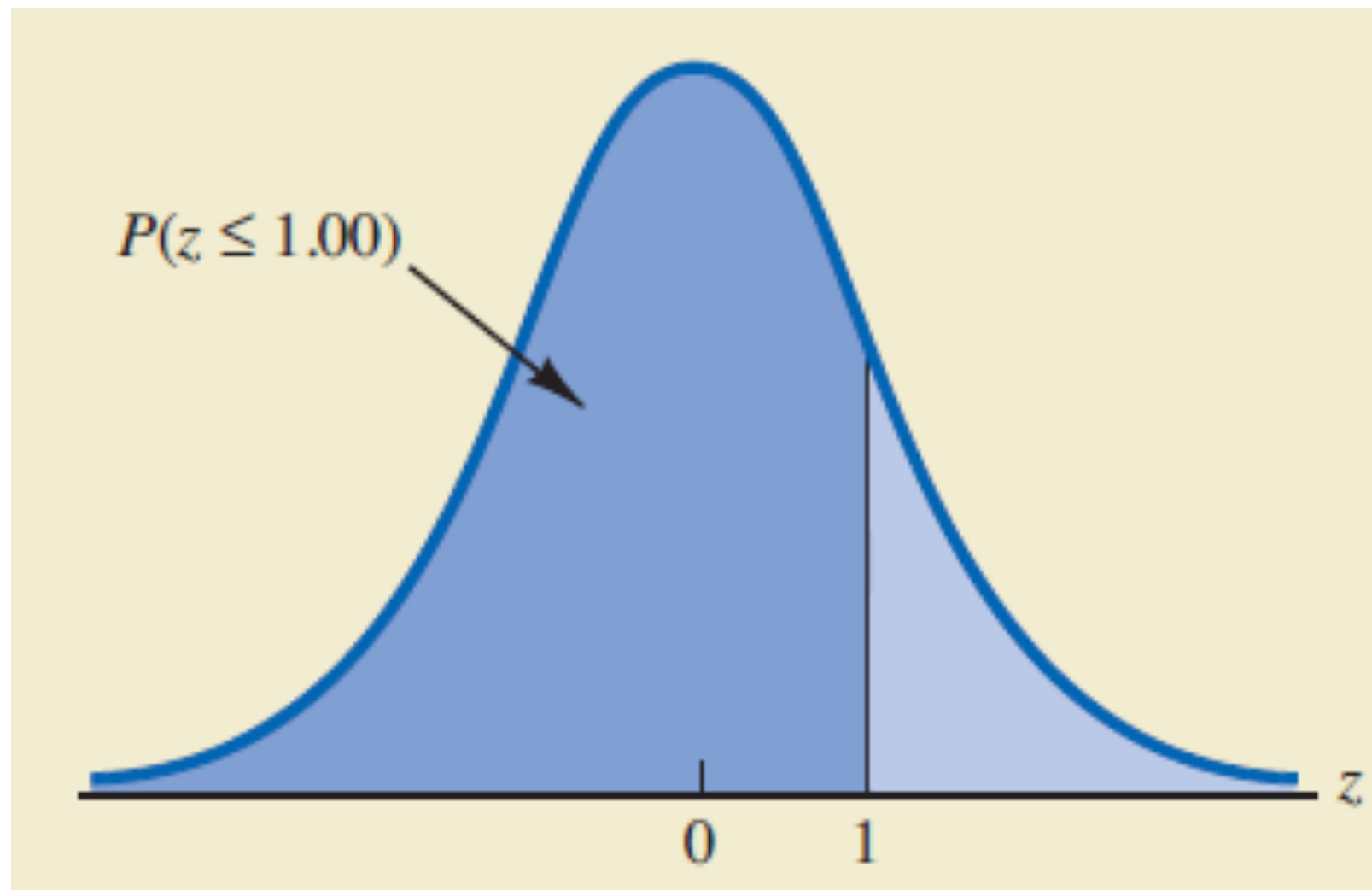
Converting to the Standard Normal Distribution

$$z = \frac{x - \mu}{\sigma}$$

We can think of z as a measure of the number of standard deviations x is from μ .

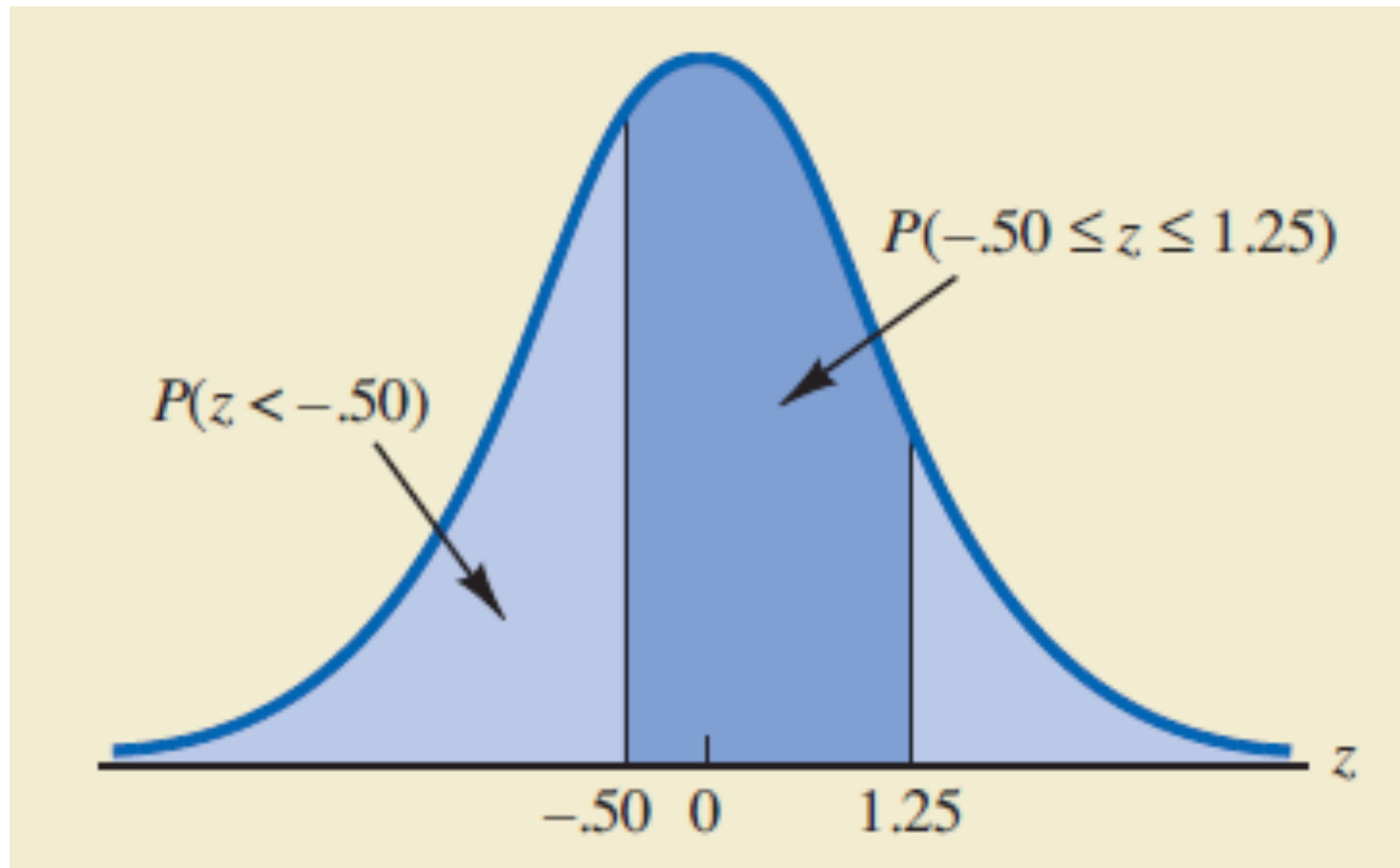
正态分布

- 问题1： z 小于等于 1.00 的概率是多少？



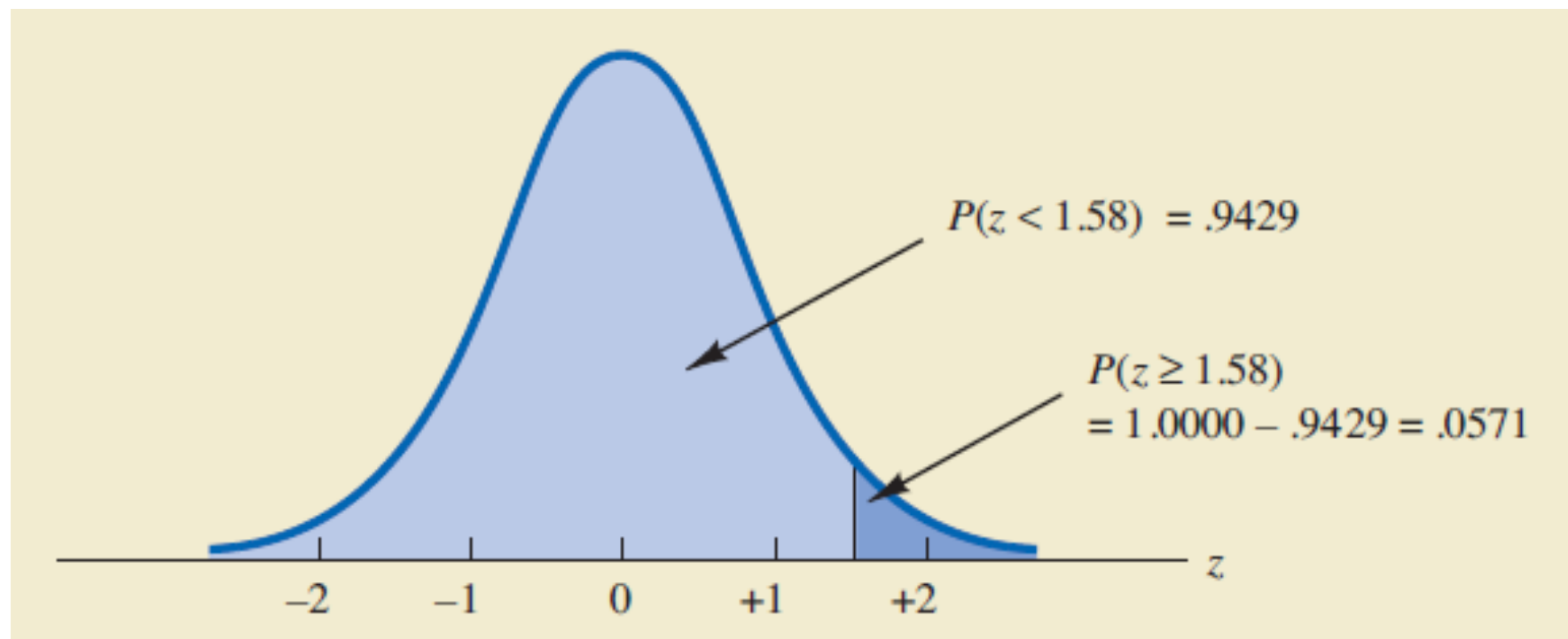
正态分布

- 问题2: z 在-0.5到1.25之间的概率? $P(-.50 \leq z \leq 1.25)$.



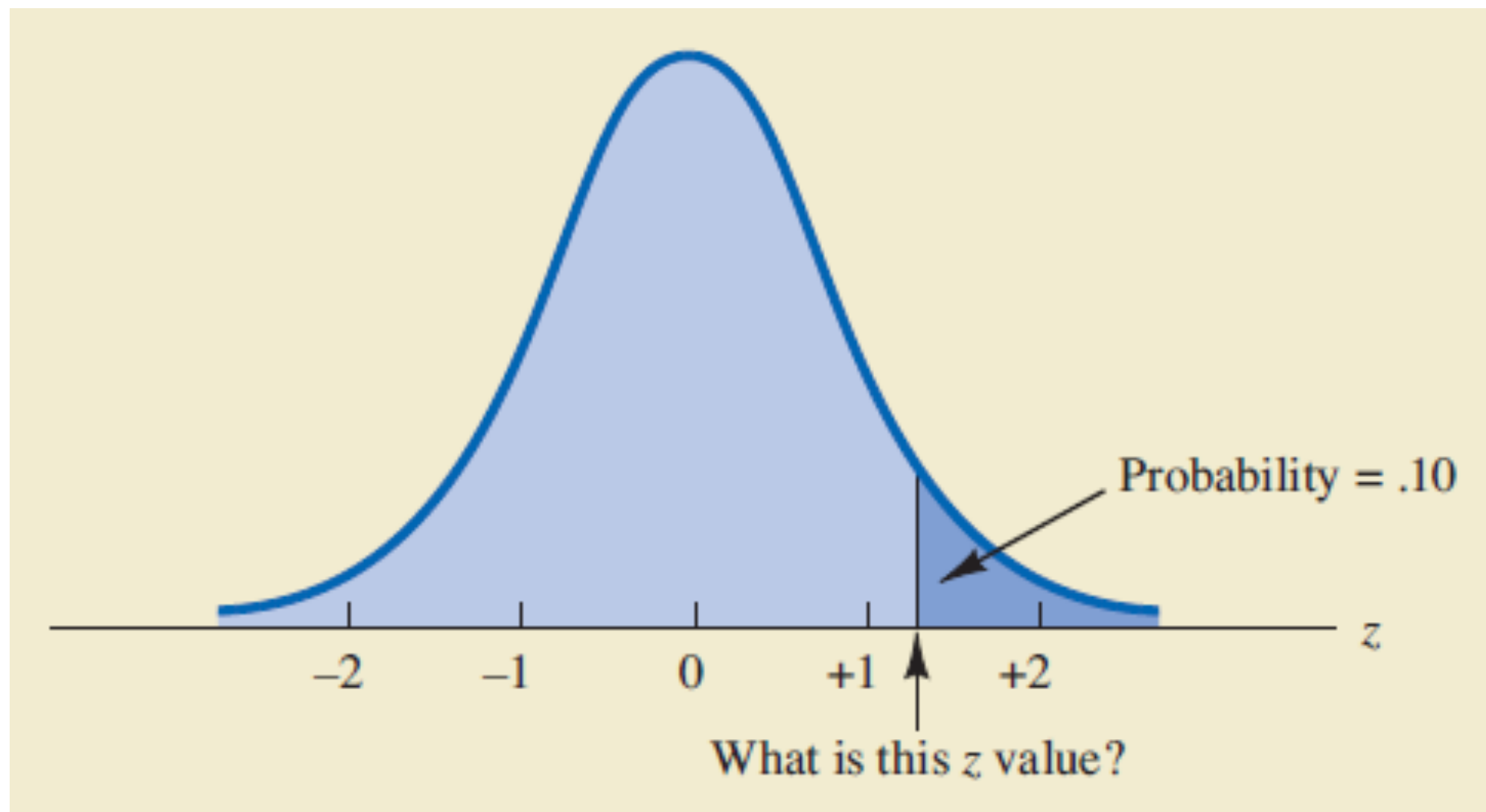
正态分布

- 问题3：z大于1.58的概率？



正态分布

- 问题4: Find a z value that the probability of obtaining a larger z value is 0.10



正态分布

- 练习:

Draw a graph for the standard normal distribution. Label the horizontal axis at values of -3 , -2 , -1 , 0 , 1 , 2 , and 3 . Then use the table of probabilities for the standard normal distribution inside the front cover of the text to compute the following probabilities.

- a. $P(z \leq 1.5)$
- b. $P(z \leq 1)$
- c. $P(1 \leq z \leq 1.5)$
- d. $P(0 < z < 2.5)$

正态分布

- 练习:

Given that z is a standard normal random variable, compute the following probabilities.

- a. $P(z \leq -1.0)$
- b. $P(z \geq -1)$
- c. $P(z \geq -1.5)$
- d. $P(-2.5 \leq z)$
- e. $P(-3 < z \leq 0)$

正态分布

- 案例：机油门店经理的决策
 - A公司是一家销售汽车机油的门店，当某款机油库存降到20升以下时，门店就会发出订货指令。
 - 经理担心的是在订货到达之前，如果出现缺货会影响门店的销售。已知在从下达订货指令到货品到达这段时间，机油的需求分布是均值为15升，标准差是6升的正态分布。
 - 经理想知道，发生缺货的概率是多少。 $P(x > 20)$.

正态分布

- 案例：机油门店经理的决策
 - 如果经理希望将缺货概率降为不超过5%。那么门店的在订货点应该设置为多少？

Wrap-up

- 抽样
- 四种概率抽样
- 抽样分布
- 正态分布