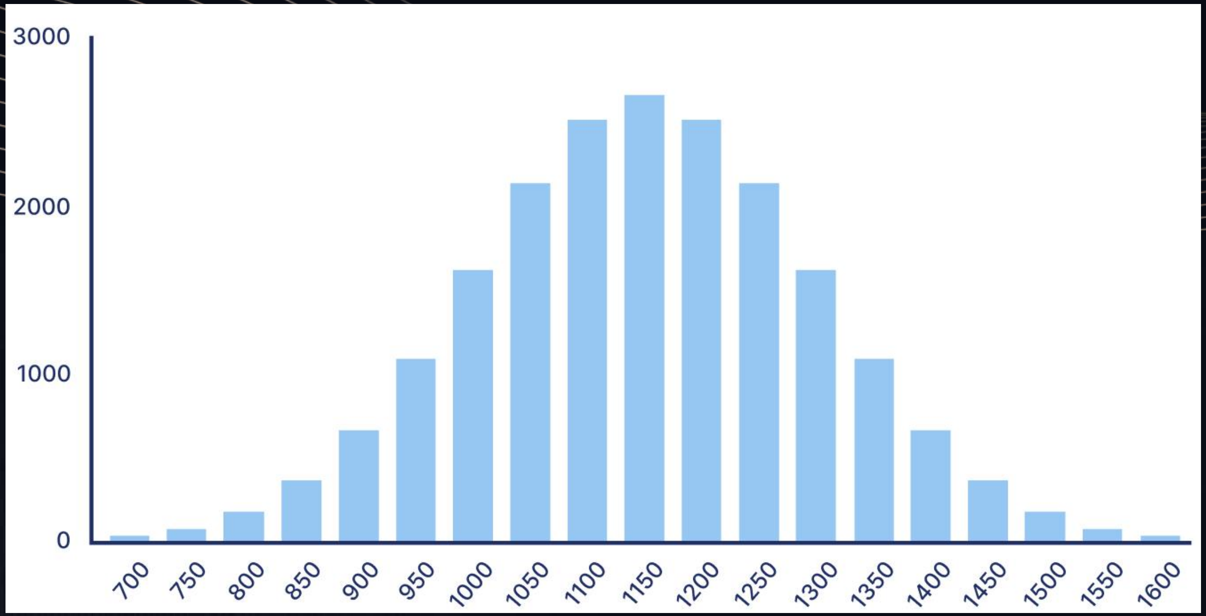
正态分布及其应用

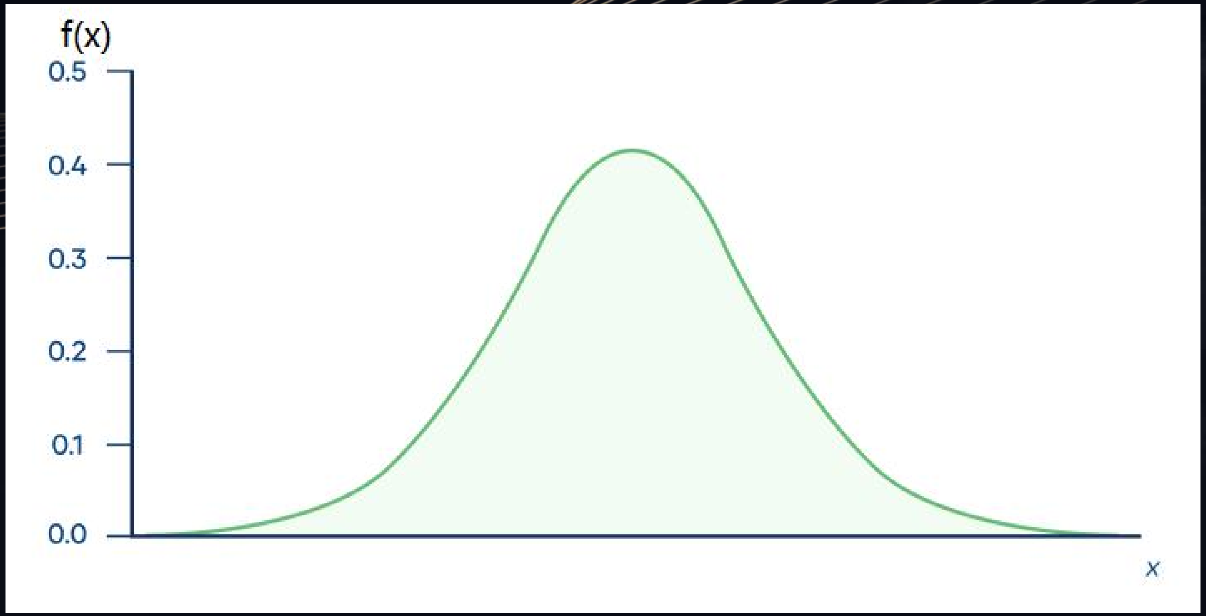
目录

- 什么是正态分布
- 正态分布的重要性
- 正态分布的突出性质
- 正态分布的特点
- 正态分布的公式和图形
- 标准正态分布的公式和图形
- 使用 z 分布查找概率
- 正态分布的经验规则
- 正态曲线下的面积
- 正态曲线下的面积规律
- 正态分布的应用

什么是正态分布

正态分布是描述连续型变量值分布的曲线，在正态分布中，数据对称分布，没有偏斜。在图形上绘制时，数据遵循钟形，大多数值围绕中心区域聚类，并在远离中心时逐渐减少。





正态分布重要吗

自然科学和社会科学中的各种变量都是正态或近似正态分布的。身高、出生体重、阅读能力、工作满意度或 SAT 分数只是此类变量的几个例子。

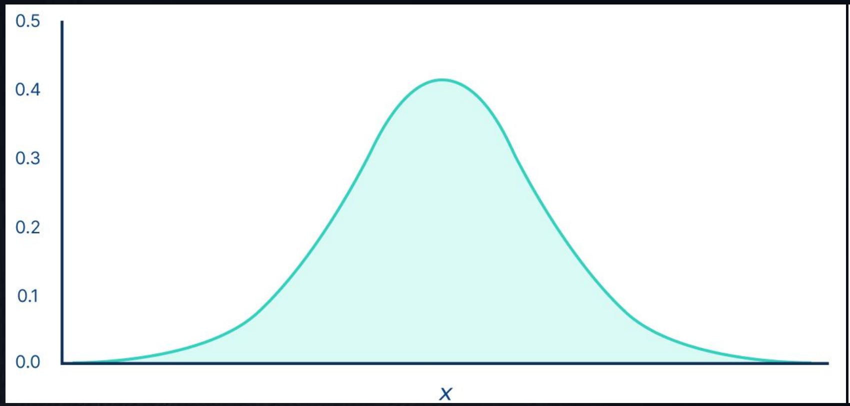
由于正态分布变量非常常见，因此许多统计检验都是针对正态分布总体设计的。

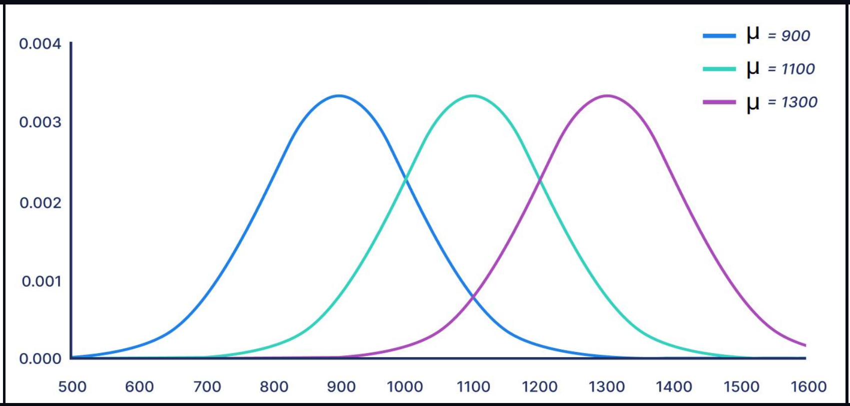
了解正态分布的属性意味着您可以使用推论统计量来比较不同的组，并使用样本对总体进行估计。

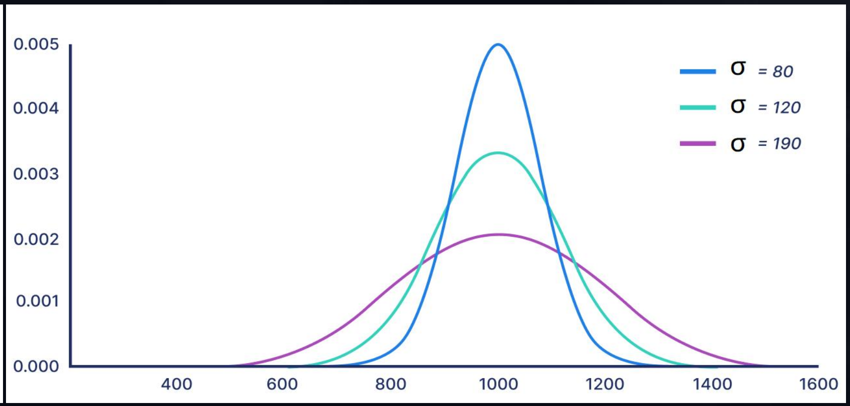
正态分布的性质是什么

正态分布的突出性质：

- 分布围绕平均值对称：一半的值低于平均值，一半高于平均值。
- 分布可以用两个值来描述：平均值和标准差。
- 平均值是位置参数，而标准差是刻度参数。
- 平均值确定曲线峰值的中心位置，增加均值使曲线向右移动，而减小均值使曲线向左移动。
- 标准差拉伸或挤压曲线。小的标准差导致窄曲线，而大的标准差导致宽曲线。

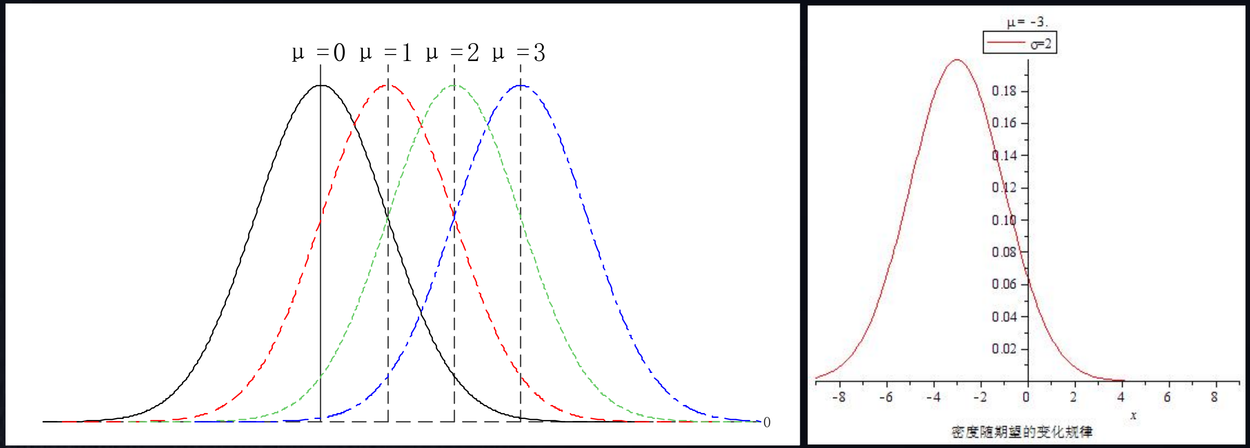




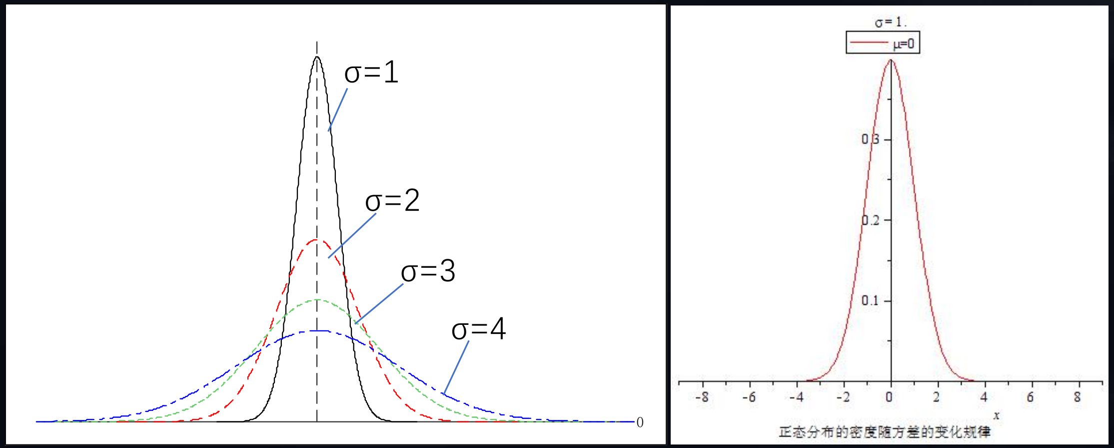


正态分布的特点

- ①曲线位于x轴上方，与x轴不相交；
- ②曲线是单峰的，它关于直线x=μ对称；
- ③曲线在x=u处达到峰值；
- ④曲线与x轴之间的面积为1，在对称轴的左右两部分完全相同，
- 即在x=μ这条直线左右两边的面积各为0.5，即$S_{(x<\mu)}=S_{(x>\mu)}=0.5$；
- ⑤当$x<\mu$时，曲线上升(增函数)；当$x>\mu$时，曲线下降(减函数)，并且
- 当曲线向左、右两边无限延伸时，以x轴为渐近线，向它无限靠近；
- ⑥当σ一定时，曲线随着μ的变化而沿x轴平移；
- ⑦当μ一定时，曲线的形状由σ确定，σ越小，曲线越尖削，表示总体的分布越集中；σ越大，曲线越平阔，表示总体的分布越分散。



σ不变，μ发生变化

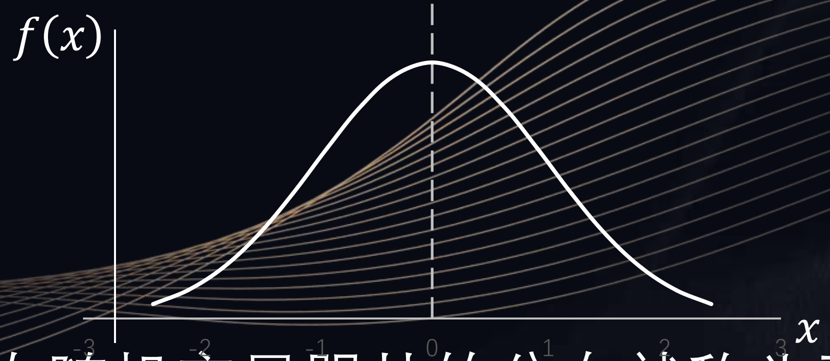


μ不变，σ发生变化

正态分布的公式和图形

正态分布是概率分布，曲线下的总面积始终为 1 或 100%，对于任何随机变量 x，将均值μ 与标准差σ代入概率密度函数公式，就可以找到该x值的概率密度$f(x)$：

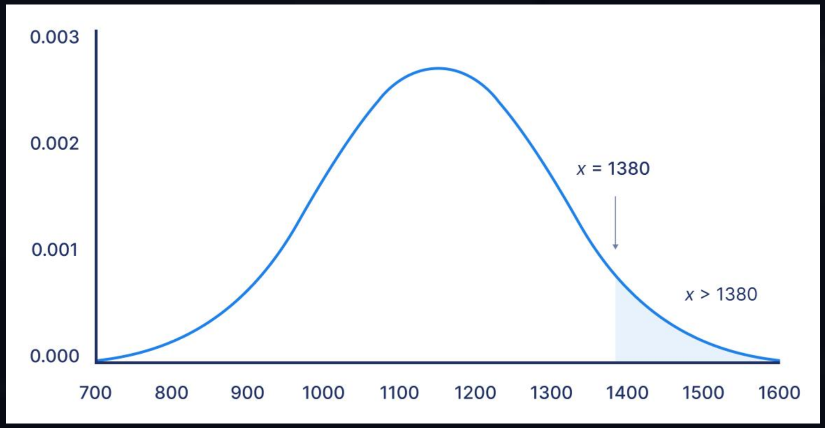
正态概率公式	解释
$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}$	$f(x)$= 概率 x = 变量的值 μ = 平均值 σ = 标准偏差 σ^2= 方差



则，这个随机变量就称为正态随机变量，正态随机变量服从的分布就称为正态分布，记作x~N（μ，σ^2），读作x服从N（μ，σ^2），或x服从正态分布。

示例1：使用概率密度函数
想知道样本中的 SAT 分数超过 1380 的概率。

如右图，在概率密度函数的图形上，概率是位于 SAT 分数等于 1380 的右侧曲线下方的阴影区域(可以使用标准正态分布获得此分数的概率值）。

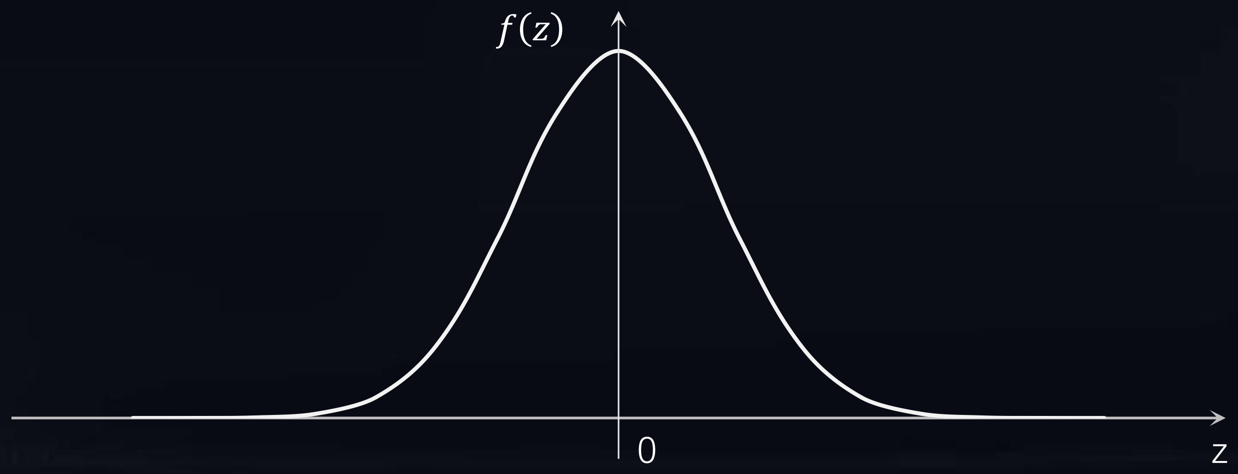


标准正态分布的公式和图形

标准正态分布（也称为 z 分布）是一种特殊正态分布，其均值μ为 0，标准差σ为 1。每个正态分布都是标准正态分布的一个版本，该正态分布被拉伸或挤压并向右或向左水平移动。

当μ=0，σ=1时，正态分布就是标准正态分布，此时概率密度函数$f(z)$为：

$f(z)$公式	z 值公式	解释
$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$	$z=\frac{x-\mu}{\sigma}$	x = 单个值 μ = 平均值 σ = 标准偏差



使用 z 分布查找概率

每个 z 值都与一个概率值相关联，该概率 值告诉我们低于该 z 值发生的可能性。如果将单个值转换为 z 值，则可以找到该值之前的所有值在正态分布中出现的概率。

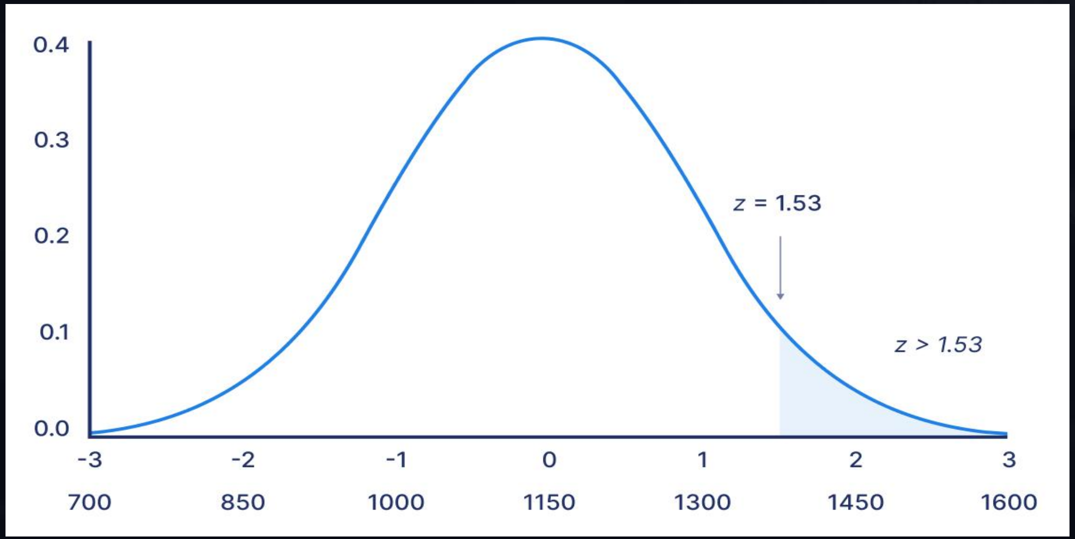
示例3：查找样本中 SAT 分数超过 1380 的概率，分布的均值为 1150，标准差为 150.
要首先得到 z 值，z 值告诉我们 1380 与平均值相差多少个标准差。

公式	计算
$z=\frac{x-\mu}{\sigma}$	$z=\frac{1380-1150}{150}=1.53$

当 z 为 1.53 时，$f(z)$ 为 0.937，这是 SAT 分数为 1380 或更低的概率，要获得阴影区域的概率(面积)，需要从整体中减去 0.937：

$$S_{(x>1380)} = 1 - 0.937 = 0.063$$

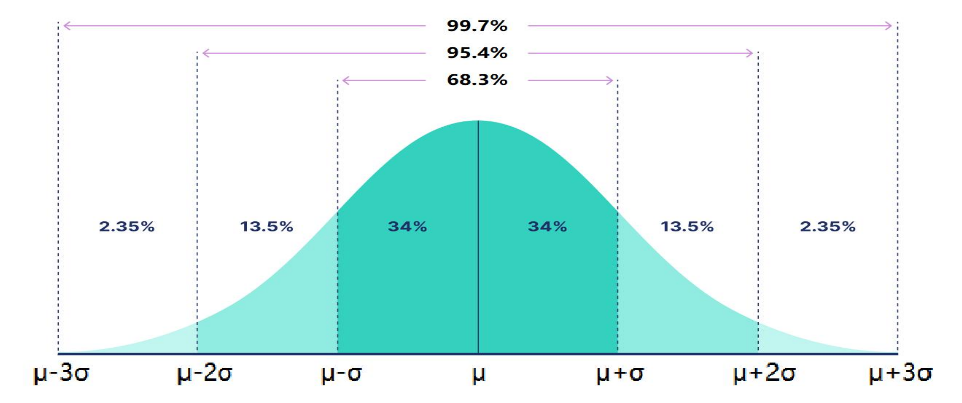
也就是，样本中只有 6.3% 的 SAT 分数超过 1380。



正态分布的经验规则

3σ规则告诉我们大多数值在正态分布中的位置：

- 大约 68.3% 的值在平均值的 1 个标准差以内;
- 大约 95.4% 的值在平均值的 2 个标准差以内;
- 大约99.7% 的值在平均值的 3 个标准差以内.

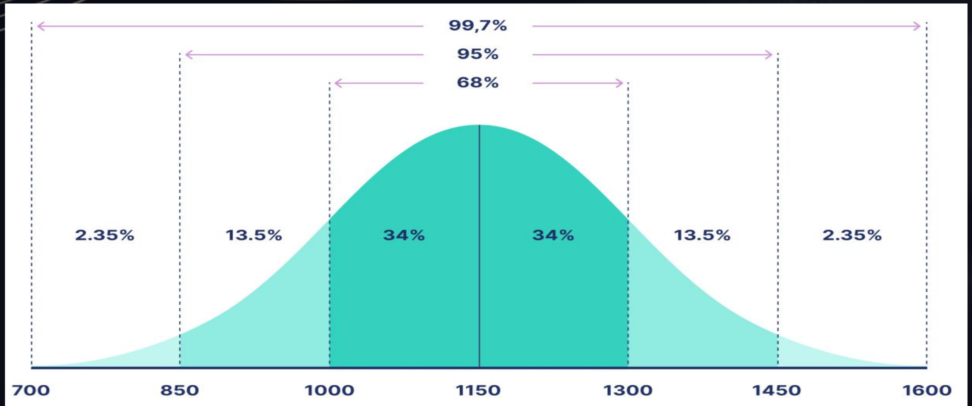


在正态分布中使用经验规则：

示例2：您在新的考试准备课程中收集学生的 SAT 分数。数据服从正态分布，平均得分μ为 1150，标准差σ为 150。

- 遵循经验规则:

大约 68% 的分数在 1000 到 1300 之间，高于和低于平均值 1 个标准差。
大约 95% 的分数在 850 到 1450 之间，高于和低于平均值 2 个标准差。
大约 99.7% 的分数在 700 到 1600 之间，高于和低于平均值 3 个标准差。



经验规则是一种快速方法，用于获取数据的概览，并检查是否存在任何不遵循此模式的异常值或极值。

正态曲线下的面积计算

正态曲线下面积:

图2中阴影部分（-∞，x_1）的面积称为正态分布的分布函数，记为:

$$f(x)=\int_{-\infty}^{x_1}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\mathrm{d}x$$

图4中阴影部分（代表任意区间）的面积计算:

$$f(x_2)-f(x_1)=\int_a^b\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\mathrm{d}x$$

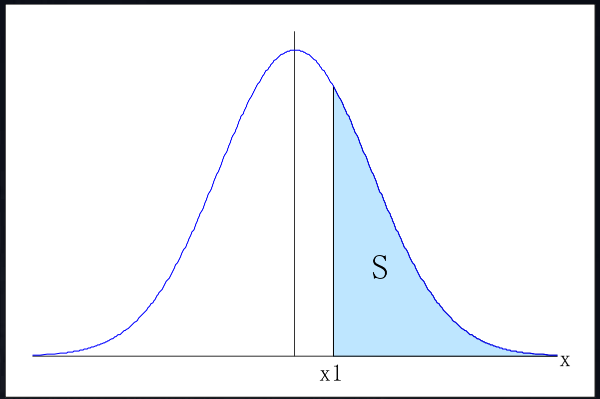


图1

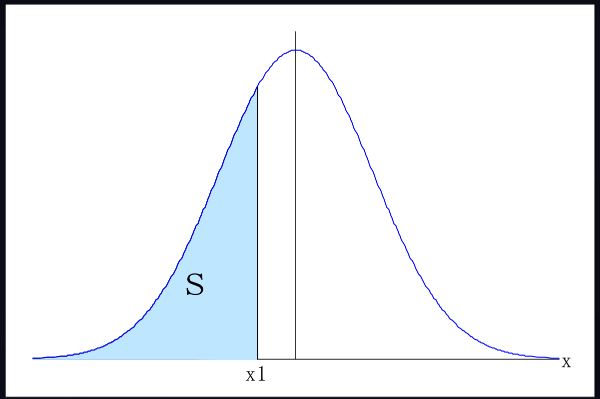


图2

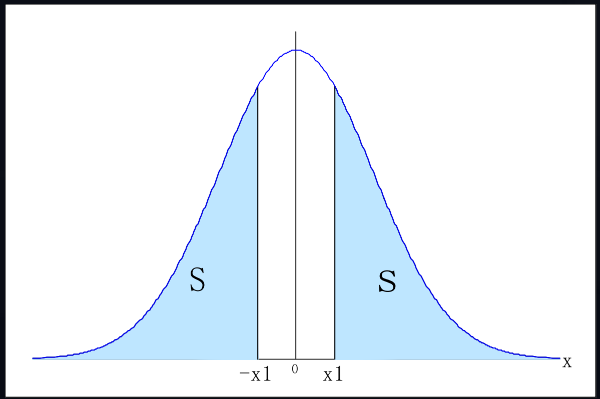


图3

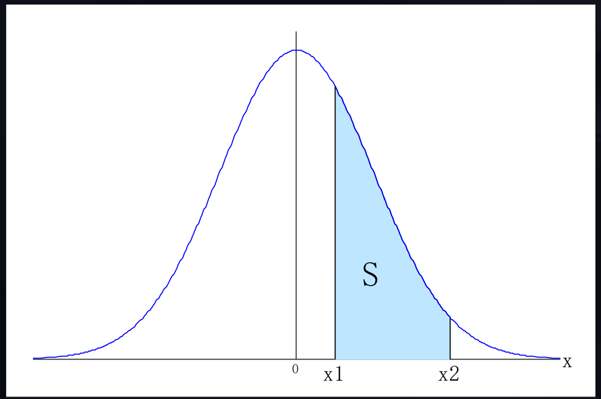


图4

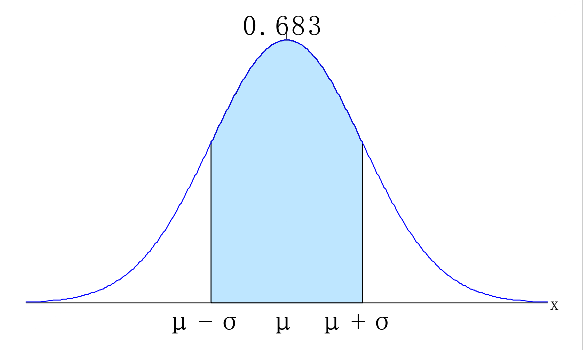
正态曲线下的面积规律

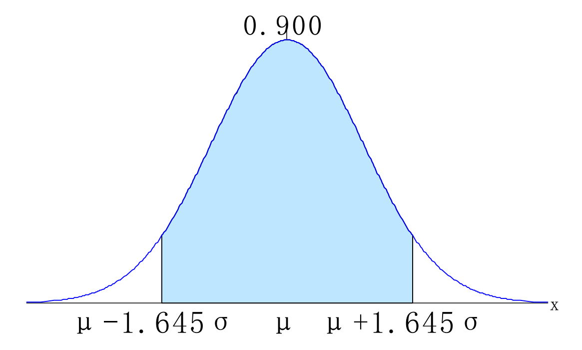
常用的正态曲线下面积及其对应的分位数:

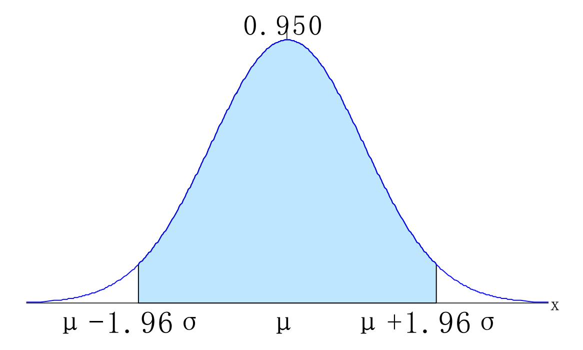
表1：正态分布曲线下的面积及其分位数

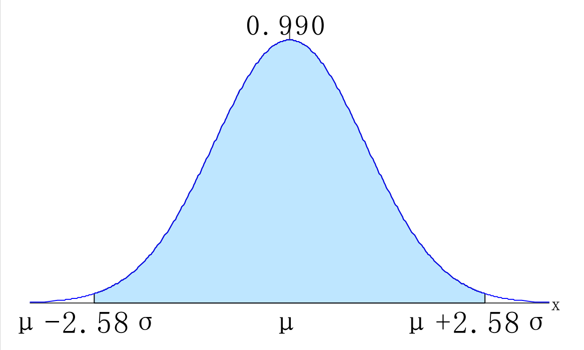
标准正态分布N（0,1）	一般正态分布N(μ，σ²)	面积（%）
−1<u<1	−σ<x<σ	68.27
−1.645<u<1.645	μ−1.645σ<x<μ+1.645σ	90.00
−1.96<u<1.96	μ−1.96σ<u<μ+1.96σ	95.00
−2.326<u<2.326	μ−2.326σ<u<μ+2.326σ	98.00
−2.58<u<2.58	μ−2.58<u<μ+2.58σ	99.00

$$\mu=\frac{x-\mu}{\sigma} \Longrightarrow x=\mu+u\sigma$$









正态分布的应用

示例4：学校 95% 的学生身高在 1.1 米到 1.7 米之间，假设这些数据是正态分布的，计算平均值和标准差。

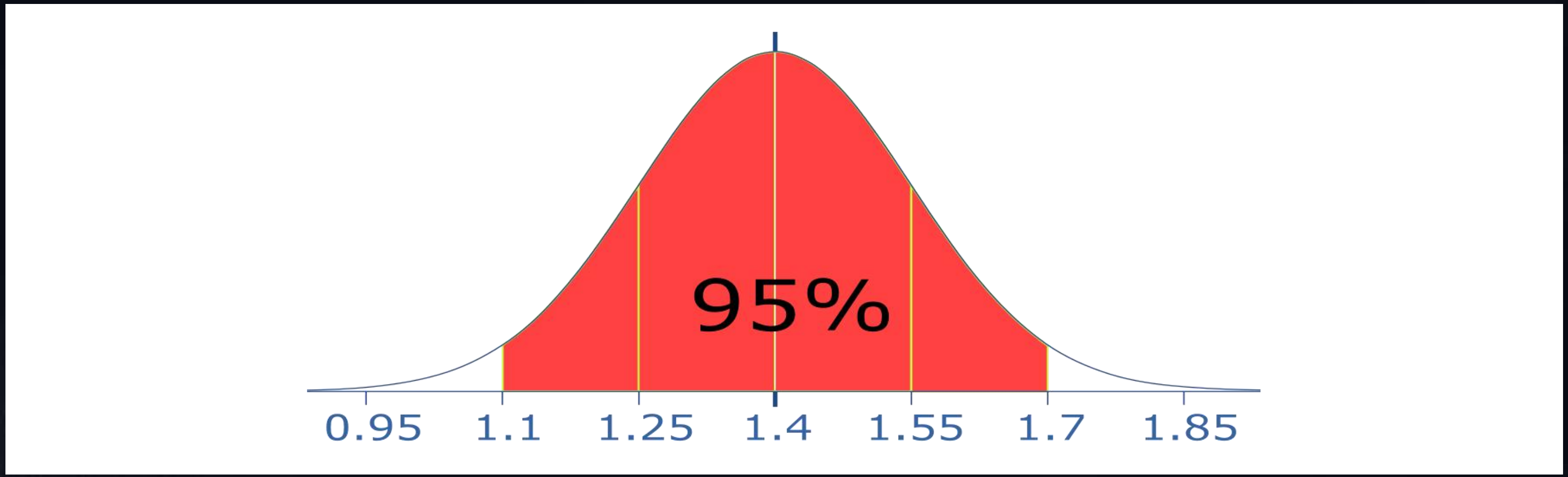
平均值 = （1.1m + 1.7m） / 2 = 1.4m

95% 是平均值两侧的 2 个标准差（总共 4 个标准差），

故：1 个标准差 = （1.7米-1.1米） / 4

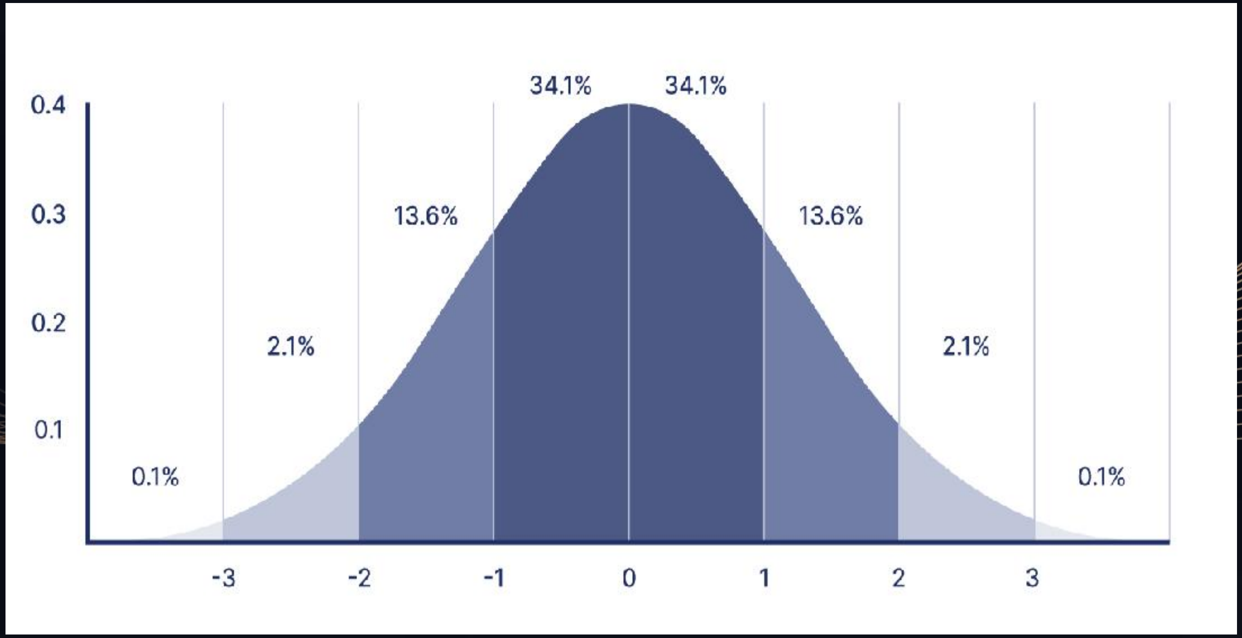
= 0.6米 / 4

= 0.15米



正态分布的应用

95%的交易仓单都是微赚微亏的状态，但是这95%的交易仓单还必须交易，不能不交易去回避，如果可以回避这95%的噪音仓单，那么正态分布就不会是客观规律了，这95%的交易仓单是不以人的意志为转移必须的操作，比较任何个人之间或自己之间那95%的交易记录，是看不出来水平的。能够体现真实水平的是那95%以外的交易部分，因为真正盈利或亏损是的仓单也就是那么2%的交易次数导致的，大赚靠那2%，大亏也是靠那2%；99%的人亏损是因为那2%的大赚机会错失了，而那2%的大亏全部抓住了；1%的人盈利是因为那2%的大赚机会抓住了，而那2%的大亏错失了，从而导致了1%与99%的天囊之别，关键那几次搞错了，关键时刻掉链子，这是交易结果差距所在。



个人情绪价值是常态分布，由所有市场参与者个人情绪价值构成的指数或者价格形成的群体情绪价值，当然也符合常态分布。那么这种常态分布，不是体现在个人仓单交易记录上，而是体现在价格涨跌幅的分布上。其中boll带所揭示的正是常态分布标准差模式下的概率涨跌幅度，boll带的上下轨并没有支撑、阻力作用，这是很多人犯的错误，总把boll上下轨当支撑、阻力而误入歧途。Boll带上下轨反应的是标准差倍数，中轨反应的是均值回归，2倍boll带的上轨到中轨或中轨到下轨，表明当周期下价格有95%的概率落在这个振幅内，1倍boll的上轨到中轨或中轨到下轨，表明当周期下价格有68%的概率落在这个振幅内。

无论是个人操作记录还是价格指数的运行符合正态分布，那么对于周期范围内的操作，95%的时间都是持仓状态，这95%的时间绰绰有余为交易进行思考，5%的时间是群体贪婪或恐慌状态，也就为反向操作奠定了基础。连续几周暴涨要清楚一个季度的上涨大概率结束而不能追涨，连续几周的暴跌要清楚一个季度的下跌大概率结束而不能杀跌，连续三五个月的上涨要清楚未来一两年的上涨结束而应该平多单，连续三五个月的下跌要清楚未来一两年的下跌结束而应该平空单。