MM5425 商业分析

WEEK 2 LECTURE – SAMPLING

第二课 内容

- 描述性统计 Descriptive Statistics
- 抽样与估计 Sampling and Estimation

抽样性统计

SHOW AND TELL

数据

- 结构化数据:

No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000

- 非结构化数据:
  - 电子邮件 Emails
  - 社交媒体发帖 Social media posts
  - 客户点评 Customer reviews
  - 图片和视频 Images and videos

数据的类型

- 时间序列数据：针对单一变量收集的数据，例如零售店的月销售收入。
- 名义型数据：指的是名称或类别变量。例如公司中的部门名称（人力资源、财务等）或产品类别（电子产品、服装等）。
- 有序型数据：指的是数据值来自有序集合的变量。例如员工绩效（优秀、良好、差）或客户满意度（非常满意、中立、不满意）。
- 区间型数据：指的是数据值选自区间集合的变量。例如以摄氏度测量的温度或智商（IQ）得分就是区间型量表的例子。
- 其他类型数据

销售分析

当你将要讨论包含公司每日交易的文件时，你会怎么做?

Purchase Date	Customer ID	Gender	Marital Status	Homeowner	Children	Annual Income	City	State or Province	Country	Product Family	Product Department	Product Category	Units Sold	Revenue
12/18/2007	7223	F	S	Y	2	$30K - $50K	Los Angeles	CA	USA	Food	Snack Foods	Snack Foods	5	$27.38
12/20/2007	7841	M	M	Y	5	$70K - $90K	Los Angeles	CA	USA	Food	Produce	Vegetables	5	$14.90
12/21/2007	8374	F	M	N	2	$50K - $70K	Bremerton	WA	USA	Food	Snack Foods	Snack Foods	3	$5.52
12/21/2007	9619	M	M	Y	3	$30K - $50K	Portland	OR	USA	Food	Snacks	Candy	4	$4.44
12/22/2007	1900	F	S	Y	3	$130K - $150K	Beverly Hills	CA	USA	Drink	Beverages	Carbonated Beverages	4	$14.00
12/22/2007	6696	F	M	Y	3	$10K - $30K	Beverly Hills	CA	USA	Food	Deli	Side Dishes	3	$4.37
12/23/2007	9673	M	S	Y	2	$30K - $50K	Salem	OR	USA	Food	Frozen Foods	Breakfast Foods	4	$13.78
12/25/2007	354	F	M	Y	2	$150K +	Yakima	WA	USA	Food	Canned Foods	Canned Soup	6	$7.34
12/25/2007	1293	M	M	Y	3	$10K - $30K	Bellingham	WA	USA	Non-Consumable	Household	Cleaning Supplies	1	$2.41
12/25/2007	7938	M	S	N	1	$50K - $70K	San Diego	CA	USA	Non-Consumable	Health and Hygiene	Pain Relievers	2	$8.96
12/26/2007	9357	F	M	N	0	$30K - $50K	Beverly Hills	CA	USA	Food	Snack Foods	Snack Foods	3	$11.82
12/26/2007	3097	M	M	Y	1	$30K - $50K	Beverly Hills	CA	USA	Food	Baking Goods	Baking Goods	5	$14.45
12/26/2007	2741	M	S	N	3	$70K - $90K	Bellingham	WA	USA	Food	Canned Foods	Canned Tuna	4	$19.18
12/26/2007	2032	F	M	N	3	$10K - $30K	Yakima	WA	USA	Non-Consumable	Household	Plastic Products	4	$19.50
12/27/2007	6651	M	S	N	0	$30K - $50K	Portland	OR	USA	Food	Produce	Fruit	5	$13.06

如何撰写一份好的报告

- 内容 Contents
  - 统计数据 Statistics
- 演示/展示
  - 图表 Charts
- 逻辑Logic
  - PEST: 环境的政治、经济、技术和社会因素
  - 4P: 产品Product, 价格Price, 地点Place, 促销Promotion
  - 5W2H: Why, What, Who, When, Where, How, How much
  - 等等 Etc..

数值变量/定量变量的统计

- 单变量
  - 集中趋势的度量
    - 均值、中位数、众数
    - 最小值、最大值、百分位数和四分位数
  - 离散/变异程度的度量
    - 极差
    - 四分位距（IQR）
    - 方差
    - 标准差
  - 分布形态的度量
    - 偏度：当样本缺乏对称性时出现
    - 峰度：主要关注极端观测值
- 多变量
  - 协方差与相关系数

集中趋势的度量

- 均值 Mean
  - Excel function AVERAGE(区间)
  - 受异常大或异常小观测值影响 (异常值 outliers)

$$\text{Mean } (\bar{x}) = \frac{\Sigma x_i}{n_i}$$

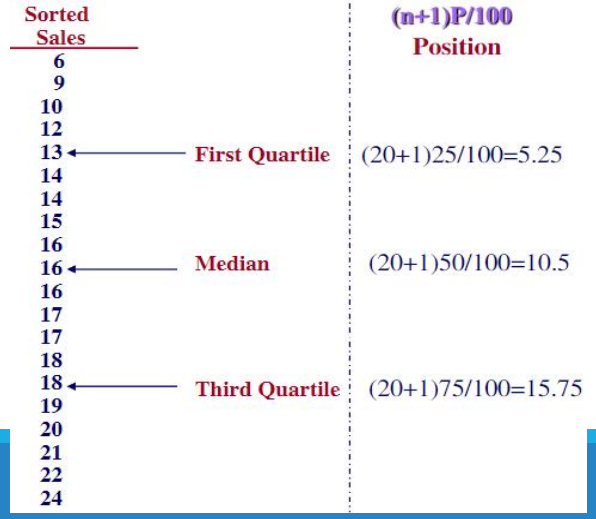
- 中位数 Median
  - 当数据按从小到大排序时，处于中间位置的数值称为**中位数**. 1, 3, 3, **6**, 7, 8, 9
  - 受极端值影响
  - Excel function MEDIAN(区间)

1, 2, 3, **4**, **5**, 6, 8, 9

- 众数 Mode
  - 出现频率最高的观测值称为**众数**
  - Excel function MODE(区间)

最小值，最大值，百分位数，四分位数

- 最小值 **Minimum** 和 最大值 **Maximum**
- 对于任意百分比 p，第 p 个百分位数（pth， percentile）是这样一个数值：有 p% 的数据小于它。
  例如，第90%分位数表示有90%的数据小于该数值。
- 四分位数将数据分为四组，每组（大约）包含全部观测值的四分之一。
  - 第一、第二和第三四分位数分别对应于 p = 25%、p = 50% 和 p = 75% 的百分位数。



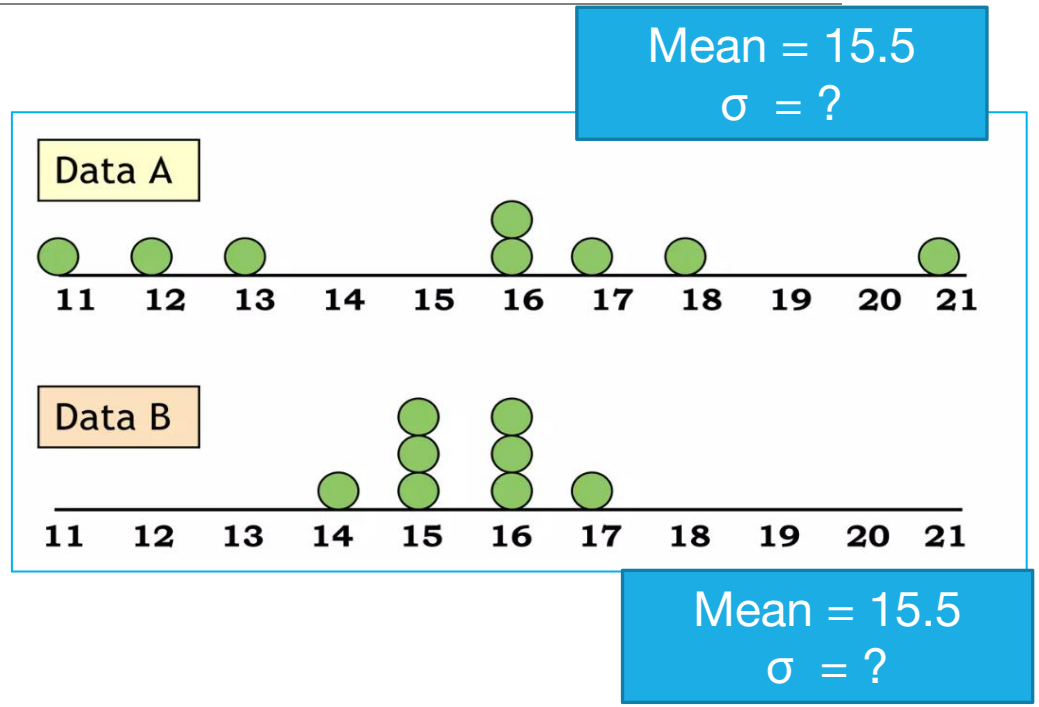
离散程度的度量

- 极差（Range）是最大值减去最小值.
- 方差是衡量数据离均值有多远的指标。它等于每个数据与均值之差的平方的平均值

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}$$

- 标准差是衡量数据离均值有多远的指标,
  - =方差的平方根: σ
- 变异系数（Coefficient of Variation, CV）
- 也称为相对标准差，是标准差与均值的比值，通常用百分数表示。它用于比较不同单位或不同均值的数据集的离散程度

变异系数（CV）=标准差/均值×100%



协方差 Covariance

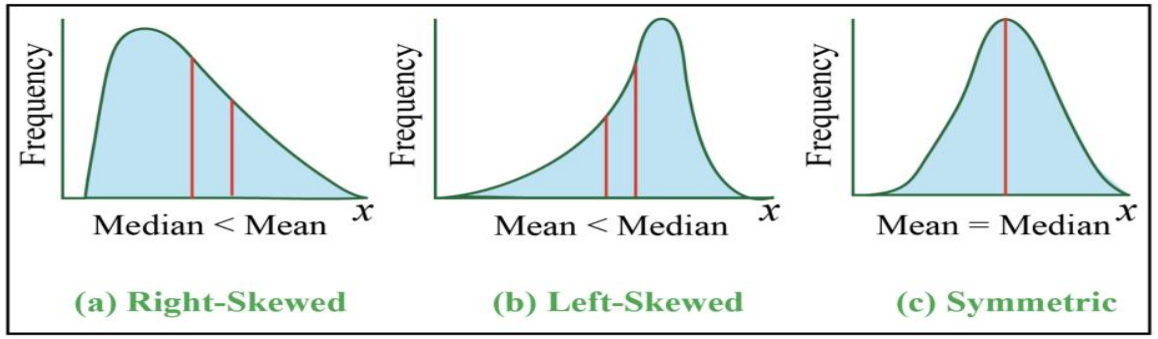
衡量两个（连续型）变量 X 和 Y 之间线性关联的指标是**相关系数**

例如, 协方差 COVARIANCE.S:

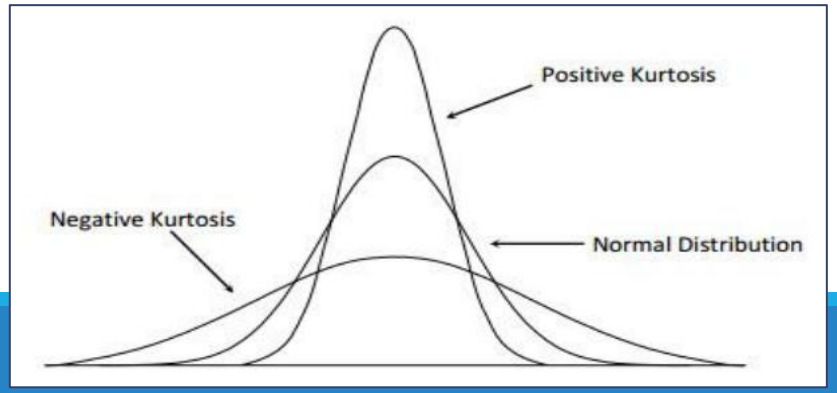
$$\operatorname{cov}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

分布形状的测量

- 偏度 Skewness
  - 测量数据集中对称分布或不对称的扭曲程度



- 峰度 Kurtosis:
  - 指分布的峰度或平坦度，峰度越低，分布越平坦



相关性 Correlation

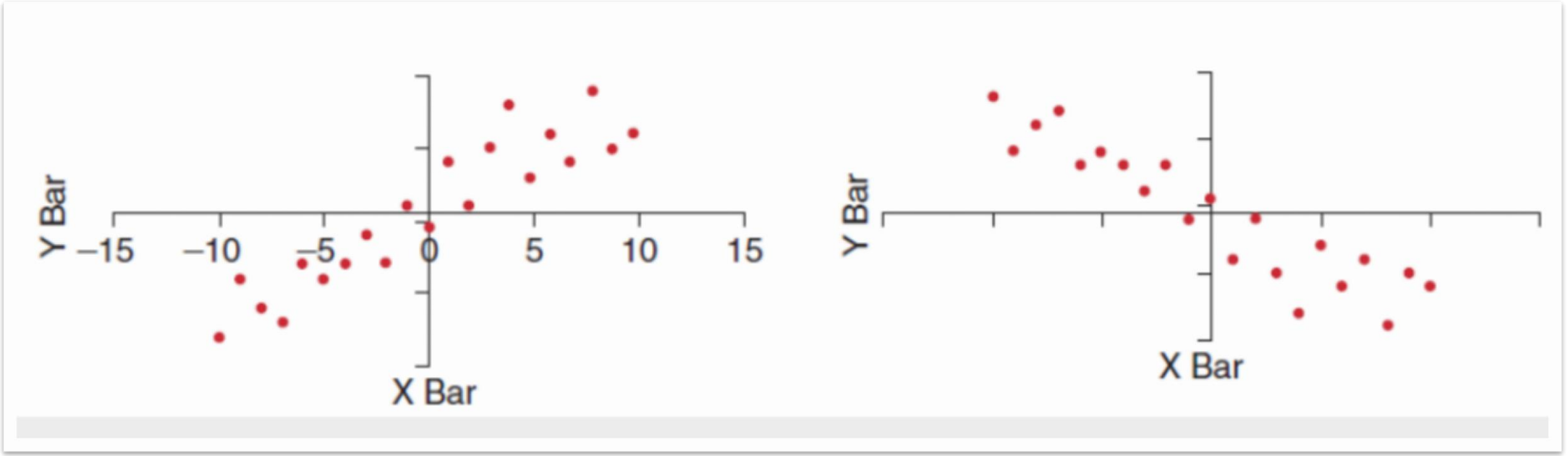
- 两个变量 X 和 Y 之间的线性关联的度量Y。
- 与协方差不同，每个变量的变化规模不会影响相关性
- 相关值始终在 -1 和 +1 之间。

$$\rho_{xy} = \frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

相关性Correlation

注意: 相关关系 ≠ 因果关系

- 我们可以看到一个地区的银行分支机构数量和总存款额呈正相关关系，这是否意味着开设更多的分支机构会带来更多的存款?



抽样与估计

SAMPLE VS POPULATION

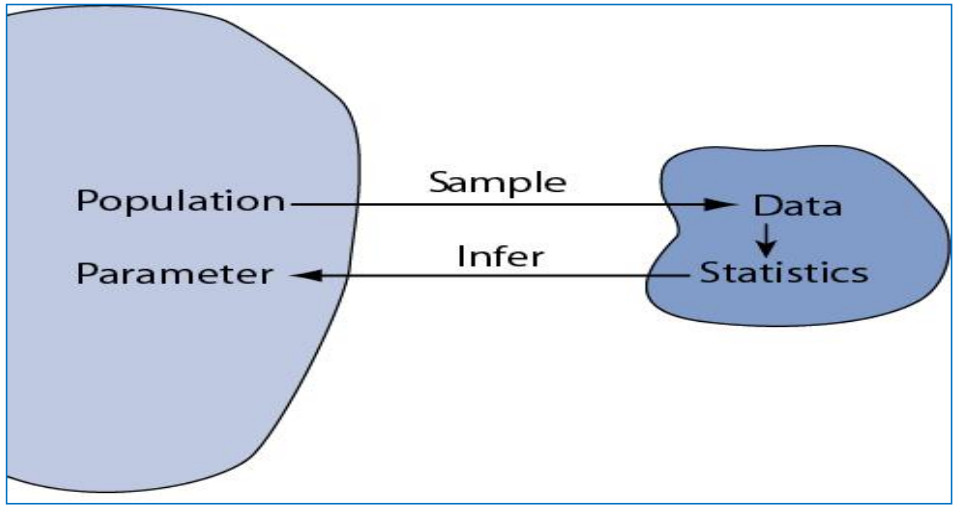
实例: 质量控制

一个大型仓库存储着数千种产品。在产品发货给客户之前，确保其质量至关重要。假设缺陷率不超过 2%

问题: 如何进行检验以确保产品质量?

取样Sampling

- 总体是所有可能观测值的集合。
- 样本是从总体中抽取的子集。
- 抽样是从总体中选择观察/记录子集的过程，以推断各种总体参数，例如平均值、比例、标准差等

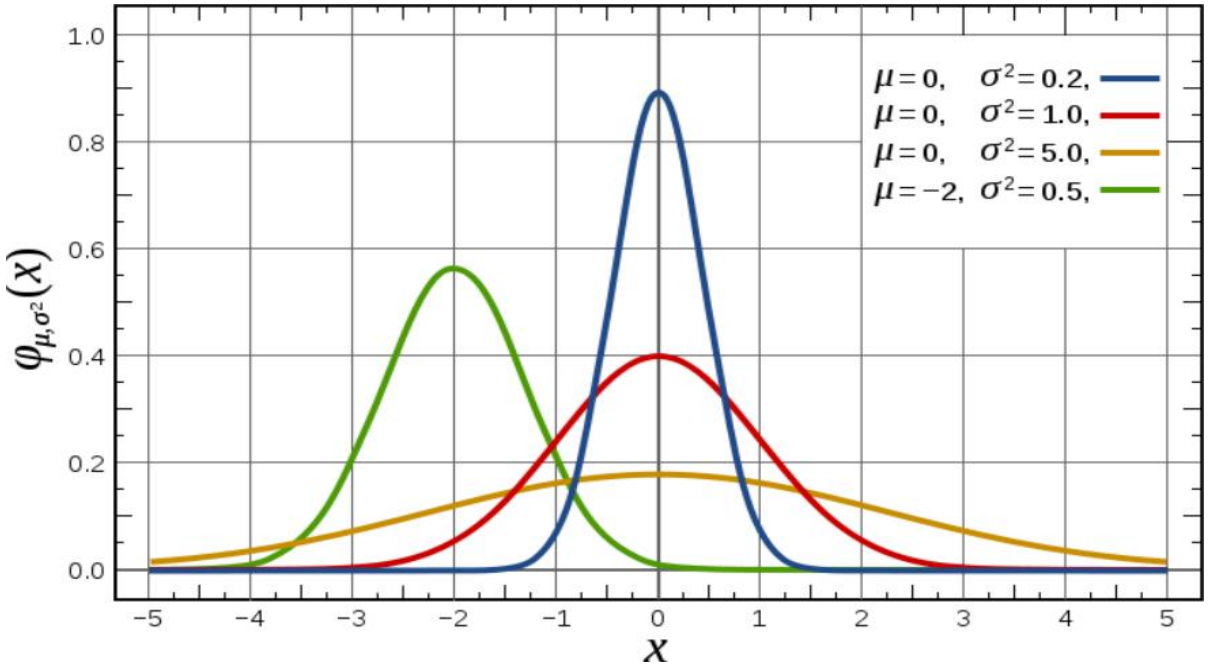


中心极限定理(CLT)

- 中心极限定理指出，对于从总体中抽取的大样本，均值的抽样分布遵循近似正态分布。
- 中心极限定理是 Z 检验和 t 检验等假设检验的基础。在很多情况下，我们只能获得一个样本，而对总体的推断必须基于样本统计量。
- CLT 的一个重要假设是随机变量必须独立且同分布。

正态分布 Normal Distribution

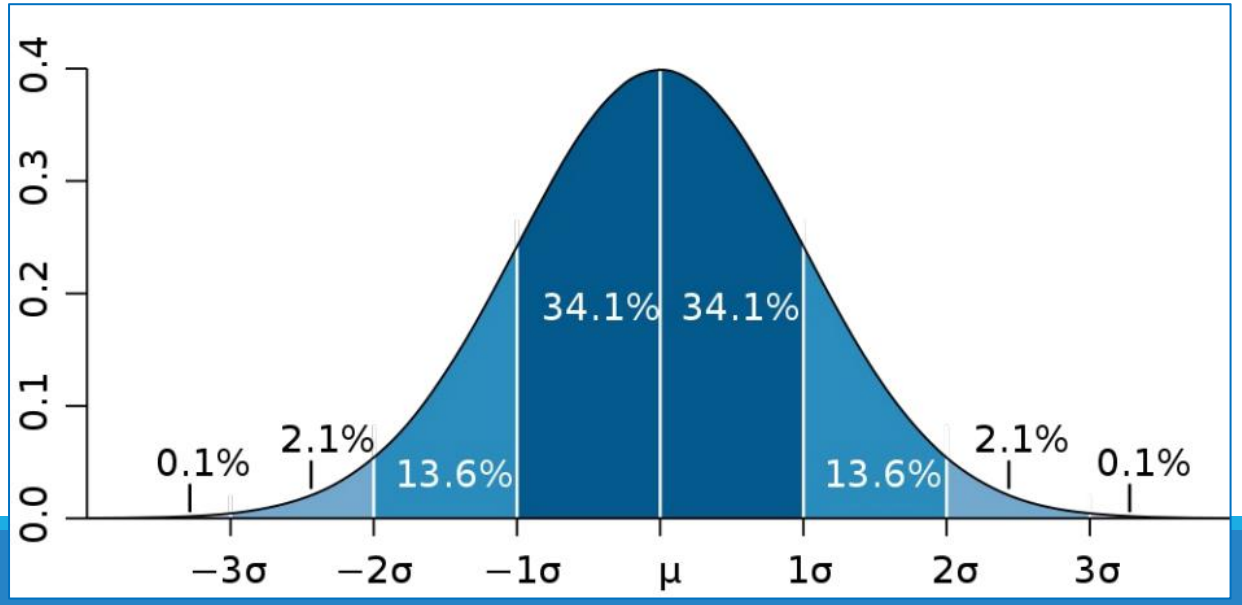
- 正态分布仅由两个参数完全表征：平均值 μ 和标准差 σ。
  - 标准正态分布的 μ = 0 和 σ = 1（图中的红线）。



正态分布曲线下的面积

6σ法则 6σ rule

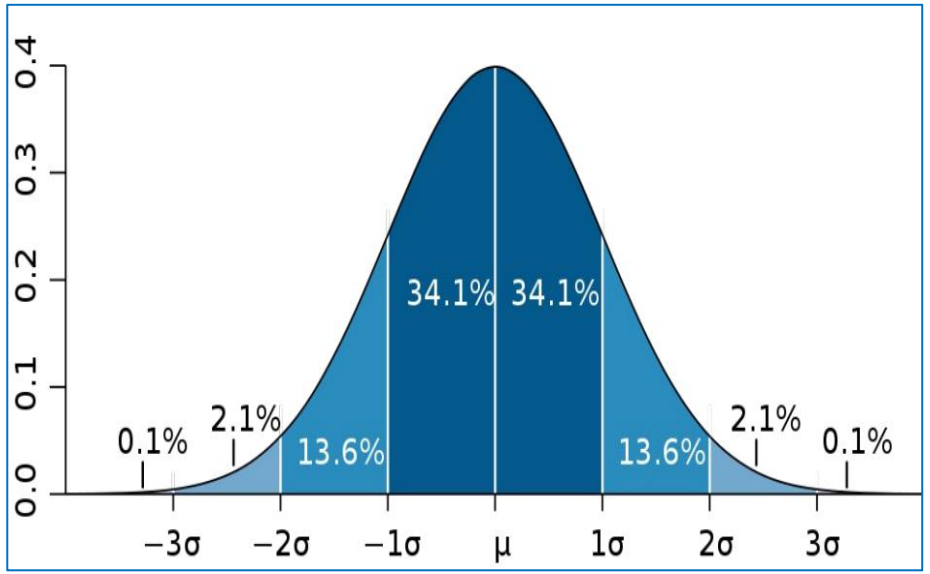
- **68%** 的数据位于均值的正负一个标准差范围内（z分数在 –1 到 1 之间）。
- **95%** 的数据位于均值的正负两个标准差范围内（z分数在 –2 到 2 之间）。
- **99.7%** 的数据位于均值的正负三个标准差范围内（z分数在 –3 到 3 之间）。



一些解释示例:

假设我们有一个正态分布，平均值为 100，标准差为 10，
我们可以推断出特定值的异常程度.

- 只有0.1%的分数高于130。
  130是均值100加上3个标准差（10 × 3），在正态分布中，超过均值3个标准差的概率约为0.1%。
- 约95%的分数在80到120之间。
  80和120分别是均值的 ± 2个标准差，正态分布中约95%的数据在这个范围内。
- 只有约5%的分数距离100超过20分。
  距离均值20分即 ± 2个标准差，约有5%的数据在这个范围 之外。
- 34%的分数在100到110之间。
  100到110是均值到+1个标准差，正态分布中约有34%的数据在这个区间。



Z检验 Z-Test

- 计算 Z 值（z 检验）的公式为:

$$\text{Z-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

其中:

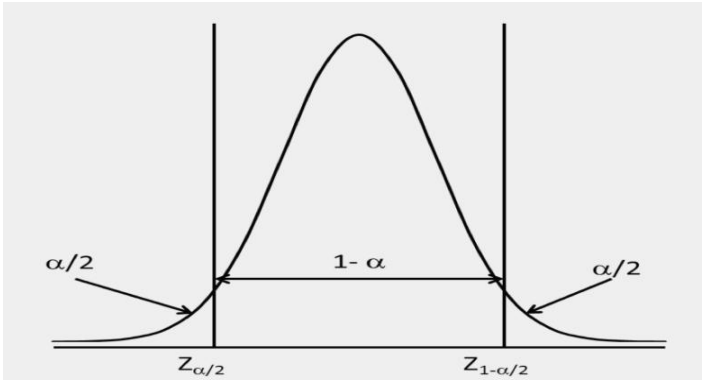
- $\bar{X}$为样本均值
- μ为总体均值
- σ为总体标准差
- n 为观测值个数

- z 检验假设检验统计量（z 分数）遵循 (0,1) 的标准正态分布

置信区间Confidence Interval (CI)

- 假设我们想要找到总体均值的 (1- α)100% 置信区间。我们可以将（在区间内未观察到真实总体均值的概率）均匀分布在分布的两侧 (α/2).
- 当总体标准差已知时，总体平均值的 (1 – a) 100% 置信区间可以推导出如下公式

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \in (-Z_{a/2},\ Z_{a/2}) \longrightarrow \boxed{(\bar{X}-Z_{a/2}*\sigma/\sqrt{n}\ ,\ \bar{X}+Z_{a/2}*\sigma/\sqrt{n}\ )}$$



常用的显著性水平值

- 总体平均值取值在和之间的概率为 (1 - α)

  介于 $\overline{X} - Z_{a/2} * \sigma/\sqrt{n}$ and $\overline{X} + Z_{a/2} * \sigma/\sqrt{n}$.

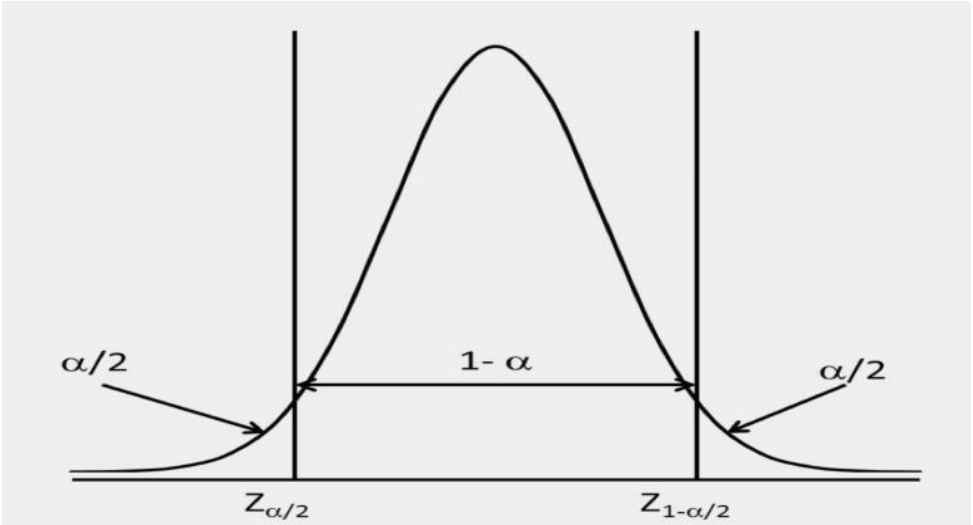
- $Z_{a/2}$ 对于不同的 α值，如下所示:

α	$\lvert Z_{\alpha/2}\rvert$	Confidence interval for population mean when population standard deviation is known
0.1	1.64	$\bar{X} \pm 1.64 \times \sigma/\sqrt{n}$
0.05	1.96	$\bar{X} \pm 1.96 \times \sigma/\sqrt{n}$
0.02	2.33	$\bar{X} \pm 2.33 \times \sigma/\sqrt{n}$
0.01	2.58	$\bar{X} \pm 2.58 \times \sigma/\sqrt{n}$

样本量估计 Sample Size Estimation

为了找到总体均值的 (1-α)100% 置信区间，样本量可以写成:

$$Z_{\alpha/2} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \rightarrow \quad n = \left(\frac{Z_{\alpha/2} * \sigma}{\bar{x} - \mu}\right)^2$$



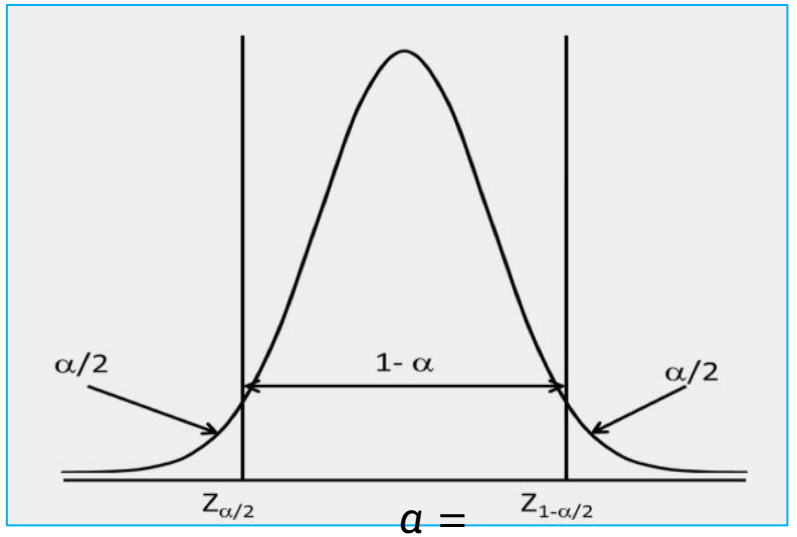
例子 1:

一家医院希望估算医生开具出院证明后，患者出院所需的时间。计算置信度为 95% 的样本量，最大估计误差为 5 分钟。假设总体标准差为 30 分钟.

解答 $\overline{X} - \mu = 5, \sigma = 30, \alpha = 0.05, |Z_{\alpha/2}| = 1.96$ for $\alpha = 0.05$

我们得到

$$n = \left(\frac{Z_{\alpha/2} * \sigma}{\overline{X} - \mu}\right)^2 = \left(\frac{1.96 * 30}{5}\right)^2 = 138$$



$\alpha = 0.05$

标准差未知时总体均值的置信区间

- 当总体平均值未知时，服从正态分布的总体平均值的 (1 - α)100% 置信区间为

$$(\overline{X} - t_{\alpha/2} * S/\sqrt{n}\,,\, \overline{X} + t_{\alpha/2} * S/\sqrt{n}\,)$$

- 在上面的等式中，值 $t_{(\alpha/2)}$，n - 1 是自由度为 (n - 1) 时 t 分布下的 t 值。
- 自由度为 (n - 1)，因为标准差是根据样本估计的

例子 2

一家在线杂货店希望估算顾客的购物篮大小（顾客订购的商品数量），以便优化用于运送杂货的板条箱的大小。

从 70 位顾客的样本来看，平均购物篮规模估计为 24，样本估计的标准差为 3.8。

计算客户订单购物篮大小的 95% 置信区间。

解答

$\overline{X} = 24,\ n = 70,\ S = 3.8,\ t_{0.025,\,69} = 1.995$

使用公式计算的购物篮子大小的置信区间为:

$$(\overline{X} - t_{a/2} * S/\sqrt{n}\ ,\ \overline{X} + t_{a/2} * S/\sqrt{n}\ ) = (24\text{-}1.995\frac{3.8}{\sqrt{70}},\ 24 + 1.995\frac{3.8}{\sqrt{70}})$$

$$= (23.09,\ 24.91)$$

Reference:
Business Statistics: A First Course, 8th edition, David M. Levine, YorkDavid F. Stephan, TechnologyKathryn A. Szabat, Pearson 2020
Business analytics: The science of data-driven decision making / U. Dinesh Kumar, New Delhi Wiley India, 2022