



# 数据分析

第19课： 我们获得的数据是什么样的？

# 描述数据，知晓现状

---

- 数据探索的第一步，往往是先了解数据的状况，这就是我们常说的描述统计

在本节课程中，我们将详细讲解这些数值：

- 数据的集中度（均值/中位数/众数）

描述了数据集中的区域

- 数据的离散度（全距/四分位数/方差/标准差）

描述了数据的稳定情况，离散度越大，说明越不稳定

- 数据之间的相关性（正相关/负相关/不相关）

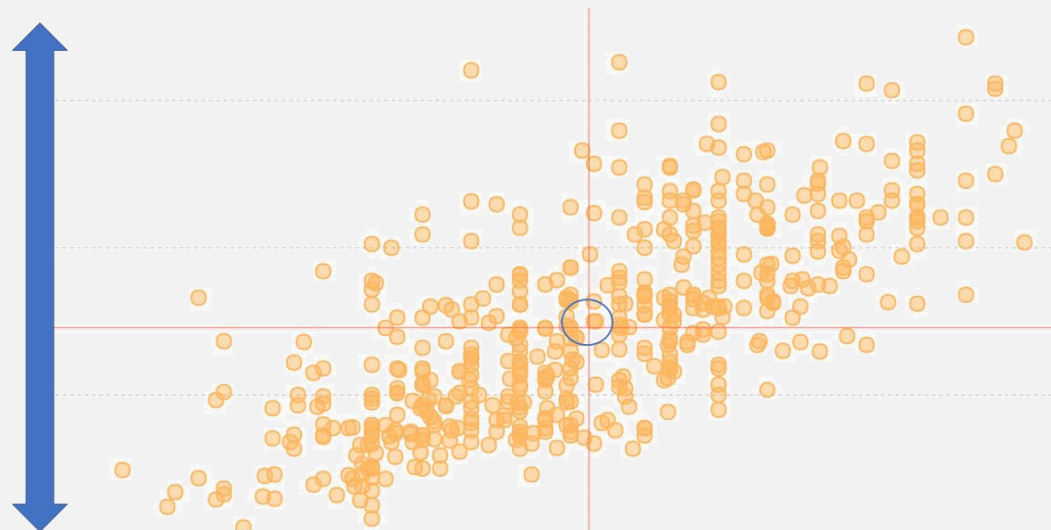
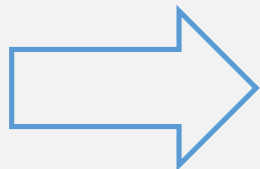
描述变量之间的关系以及关系的强弱



# 描述数据的统计学工具

**概述：**回顾前面的销售数据，我们关心的问题是，我们的销售额表现怎么样？

amount
2523.00
1097.00
1742.00
2245.00
584.00
1728.00
662.00
1438.00
2554.00
1999.00
1385.00
165.00



➤ 这里我们再引入可视化的方法来帮助理解

➤ 想象上述的数据，是散落在图上的数据点



# 描述数据的统计学工具

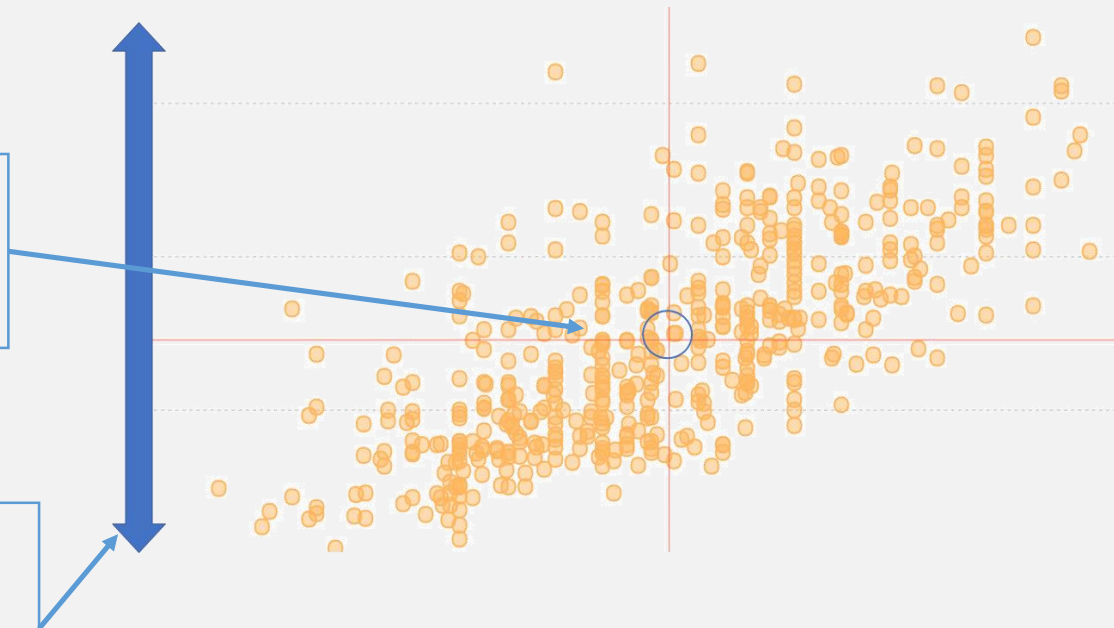
## ■ 描述数据有两个关键的问题：

### ➤ 中心度

- 图中圈的部分
- 它解释的问题是，我们的销售额的集中在哪个区域？

### ➤ 分散度

- 左侧箭头标注
- 它是最小值和最大值之间的区间
- 它解释的问题是，我们销售额有多发散，它最大和最小的点在哪里？



**描述性统计分析，就是用数字或者图表的方式来对数据进行整理和分析**



# 数据的集中度（均值，中位数，众数）

- **均值：**也称平均数，是最常用的衡量集中度的指标
  - 计算方法：所有数据的和除以数据的条数
  - 误区：极大值或极小值影响平均值的偏差

订单	1	2	3	4	5	6	7	8	9	10
销售额	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

e.g.销售额的计算结果是35K

10个数据里面8个都明显低于35K

因为平均值计算对所有数据值都“一视同仁”

所以90和95K两个非常大的销售额，一下子就拉高了整体的均值



# 数据的集中度（均值，中位数，众数）

➤ **中位数**：将数据从小到大排列之后，处于正中间位置的那个数字

订单	1	2	3	4	5	6	7	8	9	10
销售额	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

e.g.从小到大排列上述10个数字

那么 “中间 ”变成了在第5个和第6个数字之间

用15K和16K的平均值，15.5K作为中位数

比之前平均值的35K要更好的反应数据所在的 “中心”



# 数据的集中度（均值，中位数，众数）

➤ **众数**：数据中出现最频繁的那个数字

订单	1	2	3	4	5	6	7	8	9	10
销售额	12K	14K	15K	15K	15k	18K	18K	18K	90k	95k

e.g. 15K出现的频次最高，也就是我们的众数

**但因为众数只考虑频次，可能出现在一个数据中找不到众数或者有多个众数的情况**

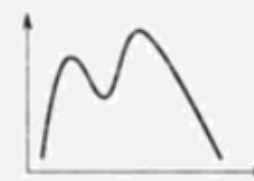
如右图例，将x轴设为数据的具体值，而y轴设为数据出现的频次



(a) 没有众数



(b) 有一个众数



(c) 有多个众数



# 数据的集中度（均值，中位数，众数）

## ■ 只是用中心度来描述数据是有缺陷的

- 大量数据汇聚在一个数字上，丧失掉了很多重要信息
- 两个不同的数据，有可能在上述指标中变成完全相同的数据

A组	25	27	29	32	33	35	37
B组	25	26	27	32	35	36	37

e.g.这两组用户的年龄，平均值都是31.1，中位数都是32，没有众数（因为所有年龄都只出现1次）  
如果只看中心度的指标，这两组是“一样的”数据，但事实呢，明显他们不是





# 数据的离散度（全距，四分位数，方差和标准差）

➤ **全距**：数据中最大值和最小值的差，是简单实用的指标

- 计算方式：在衡量离散度中，全距也就是数据中最大值和最小值的差，也是简单实用的指标
- 作用：它能用来说明数据中变大的范围
- 局限性：只用了最大值和最小值两个数据点，因此这种衡量只能提供较为粗略的信息

A组	25	27	29	32	33	35	37
B组	25	26	27	32	35	36	37

e.g.再看下我们前面所用到的A组-B组用户的数据，你会发现他们的全距也是一样的，都是12

全距计算：37-25=12



# 数据的离散度（全距，四分位数，方差和标准差）

## ➤ 四分位数：不仅使用了最大值和最小值，更考虑到了数据内部的变化

- 计算方式：将数据从小到大排序将数据“均分为”四等分的数值



**第一四分位数 (Q1)**，又称“**较小四分位数**”，等于该样本中所有数值由小到大排列后第**25%**的数字

**第二四分位数 (Q2)**，又称“**中位数**”，等于该样本中所有数值由小到大排列后第**50%**的数字

**第三四分位数 (Q3)**，又称“**较大四分位数**”，等于该样本中所有数值由小到大排列后第**75%**的数字

确定四分位数的**位置**公式如下：

- $Q1\text{的位置} = (n+1) \times 0.25$
- $Q2\text{的位置} = (n+1) \times 0.5$
- $Q3\text{的位置} = (n+1) \times 0.75$

A组	25	27	29	32	33	35	37
----	----	----	----	----	----	----	----

使用上述的公式，我们可以计算找到三个四分位数分别是标记的三个数：

$$Q1\text{的位置} = (7+1) \times 0.25 = 2$$

$$Q2\text{的位置} = (7+1) \times 0.5 = 4$$

$$Q3\text{的位置} = (7+1) \times 0.75 = 6$$



# 数据的离散度（全距，四分位数，方差和标准差）

## ➤ 方差：更全面的反应数据的离散度

描述的是数据中每个数据点和平均值偏离的距离

总体方差计算公式:  $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$

实际工作中，总体均数难以得到时，应用样本统计量代替总体参数，经校正后，样本方差计算公式为：

具体计算公式: 
$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$\bar{x}$ ：数据的均值（均为31.1）

n：样本量（均为7）

- 用每个数字减去均值，然后求这些结果的平方和
- 方差的核心概念就在这个相减再求平方和的过程中

## ➤ 标准差：方差的开方结果

- 第二个衡量分散度的指标，标准差实际上就是方差的开方结果
- 离均差平方的算术平均数的平方根，用 $\sigma$ 表示
- 标准差是方差的算术平方根

$$\sigma = \sqrt{\text{方差}}$$

↕

$$\sigma^2 = \text{方差}$$



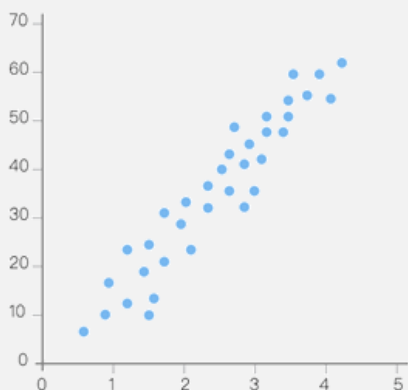
# 数据的相关性

## ■ 借助散点图理解变量间的关系

➤ 用坐标轴的x和y值分别对应需要分析的两个变量的具体数字相关关系可以分为三种可能的结果：

### 1、正相关关系

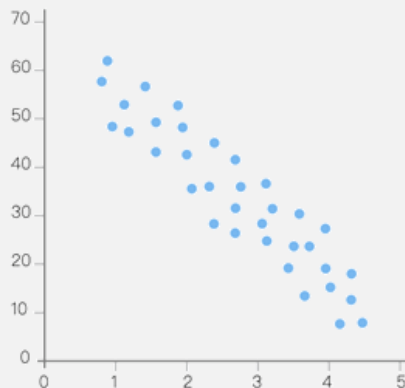
A变量增大时，B变量也随之增大



- 随着用户所停留时间增长，用户下单的平均逐步增高
- 在app上时间越长，用户的粘性越高，也更有可能会购买更多的产品

### 2、负相关关系

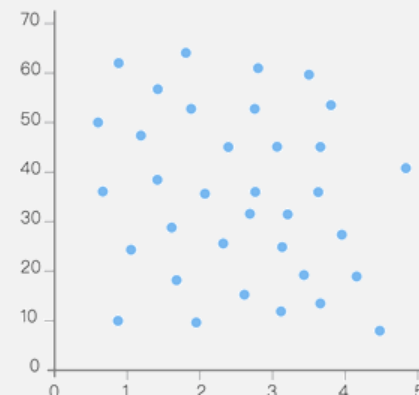
A变量增长时，B变量逐渐降低



- 个人资产越高的用户申请贷款的平均金额也是越低
- 因为那些资产高的用户，对于贷款需求本身就低

### 3、无相关关系

AB变量变化无相关



- 我们在实际中数据中也是很少见的散乱的散点图



# 数据的相关性

---

## ■ 用公式计算数据之间的相关系数：

### ➤ 相关系数的取值在-1和1之间

值越接近1：正相关关系越强

值越接近-1：负相关关系越强

值越接近0：相关关系越弱





# 数据分析

第20课： 我们关心的事件会发生吗？

# 概率与分布

---

## 概率与分布

通过对历史数据大量的重复观察，我们可以找到他们的某种规律，那种规律抽象出来，这就是我们在本节中所要学习的概率以及概率分布

## 数据类型

根据取值范围是否有限，数据可以分为类别型数据和数字型数据

## 概率的应用场景

分别对于类别型变量和数字型变量的概率和概率分布进行讲解，并具体介绍数据分析师的“万能模板”：二项式分布和正态分布



# 概率与分布

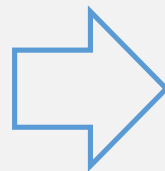
---

## ■ 对数据进行描述

数据往往有无法控制的事件组成，比如：

- 用户注册APP的时间
- 工作日用户下单购买产品的数量
- 某个产品的销售数额

通过对历史数据大量的重复观察，可以找到他们的某种规律



概率以及概况分布





# 数据类型

## ■ 两种主要的数据类型：

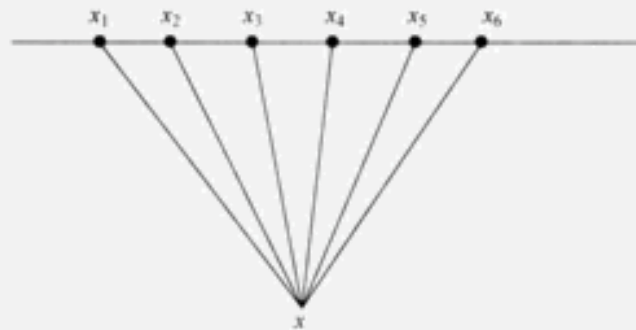
### ➤ 类别型数据

取值是有限的若干个（可以穷举出来）

如：性别，男、女

受教育程度，小学、初中、大学、硕士、博士

国内的城市，数百个



类别型数据， $x$ 只会落在某个点上

### ➤ 数字型数据

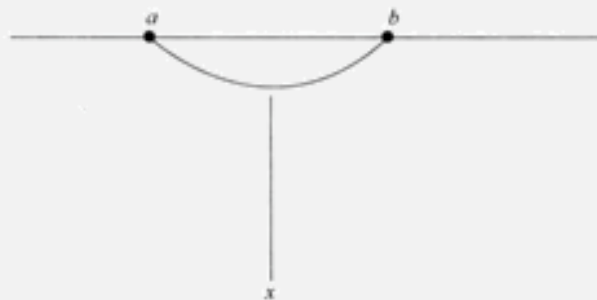
取值可能是在某个范围内的任何点（范围有时候并没有明显的界限）

如：收入的衡量

在一侧可能是以0为界限

另一侧则可以是没有限上

可以是百万可以是千万



数字型数据， $x$ 可以落在区间内的任何一个点上



# 概率的应用场景

➤ 概率：用来衡量事件发生可能性的比例

➡ 关心的时间所发生的次数除以所有发生的时间总数

应用概率方法来抽象归纳数据，对于两种不同的数据，给出的答案也会有所不同

类别型	数字型
X会有哪些值？	X会在哪个区间上取值？
(用户购买的产品种类)	(用户注册的时间在一天24小时中分布)
X取这些值的概率是多少？	X在这个区间中各个范围的概率是多少？
(购买各个产品的概率)	(在上午下午晚上的概率)

# 类别型变量的概率和分布

## ■ 应用场景

### ➤ 顾客漏斗分析

访问商品详情页的概率=商品详情页的用户数/所有浏览过商品的用户数

- 概率是60%
- 越接近1表示发生的可能性越大，而越接近0则表示越小

### ➤ 在访问商品详情页和不访问，是两个互斥的事件

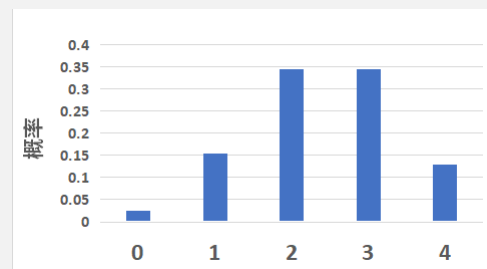
- 访问商品详情页 60%
- 未访问商品详情页 40%

### ➤ 有四个用户访问的5种可能

- 分别对应了点击的用户数从0到4
- 统计图形如右图



事件	概率
访问商品详情页	60%
未访问商品详情页	40%



# 二项式概率分布

## ➤ 只有两个结果的事件的预测：

**关注的问题：** 如果有很大的事件时，最终产生某个特定数量事件的概率是多少呢

- 比如我们的app每天都推送给50万用户消息
- 假设我们知道每个用户都有60%的概率去点击这个推送消息
- 那么最终我们获得35万用户点击这个消息的可能性是多少呢？

## ➤ 二项式分布计算公式：



$$P(x) = \frac{n!}{x! * (n - x)!} * p^x * (1 - p)^{n - x}$$



3个关键要素

- 做某件事的**总次数 (n)**
- 做这件事成功的**次数(x)**
- 这件事情成功的概率(**p=0.5**)

**我们所要计算的，就是在总次数n中，成功x次的概率P(x)**

我们还可以使用EXCEL的**BINOMDIST**函数来进行运算

=BINOMDIST(成功次数,总次数,0.5,0)



# 数字型变量的概率和分布

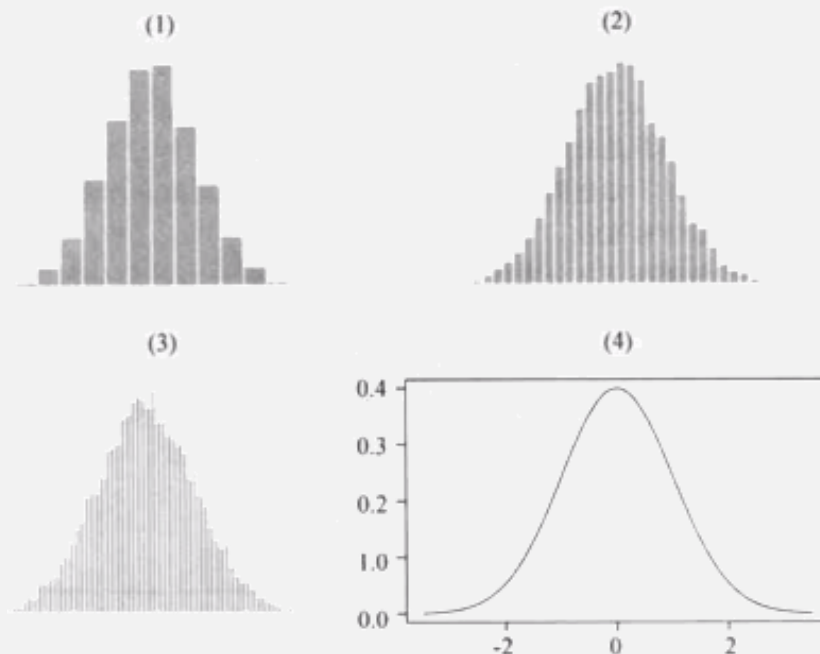
## ■ 应用场景

### 评估用户的收入水平：

- 当用户数量较小时：  
可以用分组比如0-1万，1-5万，5万-10万来描绘
- 用户数量增大：  
那么每个年龄数字都会被填充  
最终直方图也形成了连续不断的曲线

**在数字型数据中，用面积来代表了发生概率的可能性理想情况下**

- 最终成形是我们常说到的“正态分布”
- 它是数字型变量中最经典的一种概率分布



➤ 类别型数据概率直方图

直方图柱子大小代表了各个值所对应的概况情况

➤ 数字型数据中

面积代表了发生概率的可能性



# 正态分布

---



## ➤ 正态分布定义：

- 中间高两头低，左右对称，它是数字型变量中最经典的一种概率分布。
- 使用场景：身高体重、降雨量、员工绩效



## ➤ 正态分布的2个属性：

- 快速计算数据的概率分布
- 在只知道均值和方差的情况下可以知道数据的全部



# 正态分布

## ➤ 属性1 快速计算数据的概率分布

e.g.顾客的满意度打分，均值为75，标准差为6，求打分  $\leq 80$  分的概率

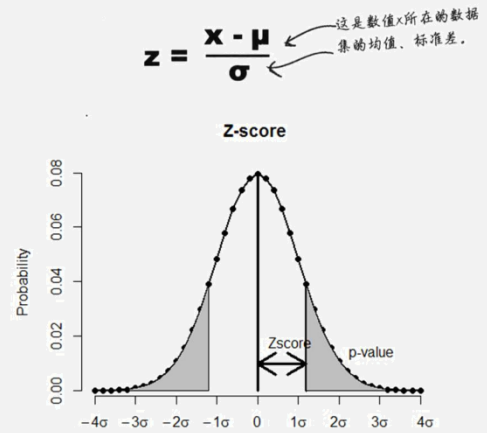
**使用Z评分：** $Z = (\text{目标数} - \text{平均数}) / \text{标准差}$

$$Z = (80 - 75) / 6 = 0.83$$

**查询Z评分概率表：**第一列中找到0.8，然后在第一行找到0.03，两者交叉的就是0.7967

顾客满意度评分小于等于80的概率是0.7967

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621



## Z标准分：快速计算数据的概率

Z标准分是一种由原始分推导出来的相对地位量数，它是用来说明原始分在所属的那批分数中的相对位置

通过计算出的Z标准分，便可以推导出对应的概率，P值

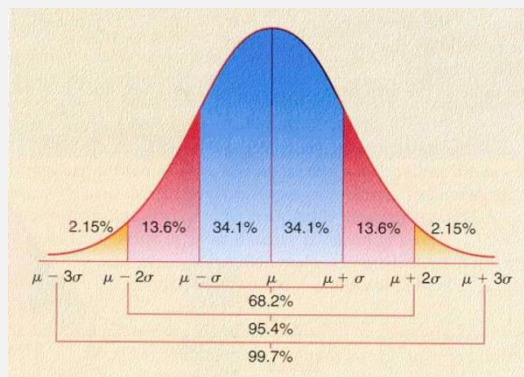
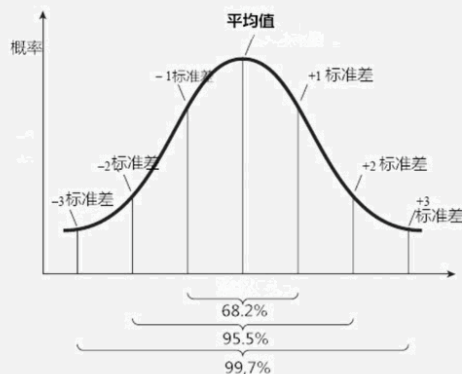
# 正态分布

➤ 属性2 在只知道均值和标准差的情况下可以知道数据的全部

- 均值 $\pm 1$ 个标准差会覆盖68.2%的数据
- 均值 $\pm 2$ 个标准差会覆盖95.5%的数据
- 均值 $\pm 3$ 个标准差会覆盖99.7%的数据
- 68.2%，95.5%，99.7%，也经常会被称作是置信水平



“在95.5%的置信水平下，均值 $\pm 2$ 个标准差会覆盖我们的样本数据”



➤ 置信区间与置信水平

- **置信区间**是指由样本统计量所构造的总体参数的**估计区间**。图中的【 $\mu - 3\sigma, \mu + 3\sigma$ 】【 $\mu - 2\sigma, \mu + 2\sigma$ 】【 $\mu - \sigma, \mu + \sigma$ 】
- 而置信区间内的数字都包含总体这一结果具有特定的概率，**这个概率就是置信水平**。图中的:68.2%、95.4%、99.7%
- **计算置信区间的方法（简易步骤）：**
  1. 求样本的平均值与标准差
  2. 确定置信水平
  3. 计算置信区间






# 概率与分布

---



在后面使用的回归或者分类分析中：

 很多都对所预测目标或者是误差等做出了特定的分析假设



通过统计软件对分布假设进行验证  
如果数据本身是不符合这些假设的  
那模型的基础就是不牢靠的





# 数据分析

第21课：假设检验，未来会发生什么？

# 抽样与估计

---

## ➤ 样本

数据和概率问题，更多的是基于所拥有的数据而来，通常称这个数据为“样本”

e.g. 初创公司，向市场上推出了新的APP

- **样本**：运行了一段时间之后，累积了一定的用户数据，这些就是“样本”
- **总体**：所有智能手机可以下载这个app的用户

## ➤ 中心极限定律

1. **样本平均值**约等于**总体平均值**
2. 不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体平均值周围，并且呈正态分布

⇒ **样本均值的方差**约等于**总体均值的方差**除以**样本数**



# 抽样与估计

## ■ 对于总体（所有智能手机可以下载这个app的用户），想了解的问题：

### 1.我们的市场用户平均每个月在我们的app下单的次数会是多少呢？

- 1万个用户抽样，发现数据的平均值为3.5
- 统计学理论证明，抽样平均数作为对总体平均值的预测
- 结论：用户每个月在我们的APP下单次数是3.5

### 2.我们的市场用户平均每个月在我们的app上花多少钱呢？

- 抽样平均数预测整体的问题 → 是一个孤立的数据点
- 实际工作中，将推测的数据放在区间内，保证预测的精确性，以及工作的灵活度
- 解决方法：使用正态分布的置信区间

抽样分组	1	2	3	4	5	6	7	8	9	10
用户花费	3000	3200	2700	3900	4000	5900	5100	4700	4000	3800

- 样本均值为3900，标准差为120
- 99.7%的数据为 $3900 \pm (3 \times 120)$  → 3540~4260
- 结论：用户平均每个月在我们的app上花钱范围为3540~4260元



# 假设验证

---

## ➤ 分析数据的思路

为了得到用户的年龄或者消费情况

参数估计，基于用户样本的数据去估计总体用户

## ➤ 假设检验的思路

根据经验或者其他方面的信息来假设一个总体用户可能的值

再根据样本情况，使用某种工具来验证这个假设是否正确

例子：

前面用户数据中平均的消费是在3900元

所有用户电商数据，“总体”的用户平均消费是在4100元

app的用户和传统的电商用户之间是否有显著的消费能力差别呢？



# 假设验证

## ■ 假设检验步骤

### 1. 设定我们的初始假设

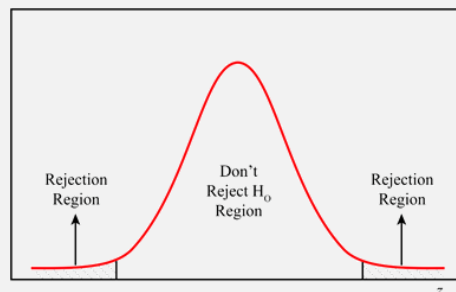
初始假设：用户平均消费是4100元

验证结果：接受或者拒绝这个假设

默认假设	对应假设
均值=4100	均值 $\neq$ 4100
均值 $\geq$ 4100	均值 $<$ 4100

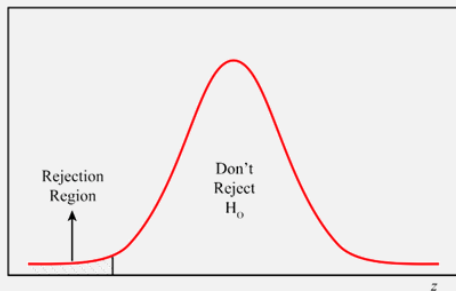
### ➤ 双尾检验:

默认假设 = , 对应假设  $>$  或  $<$   
有2个拒绝默认假设的空间



### ➤ 单尾检验:

默认假设  $\geq$  , 对应假设  $<$   
有1个拒绝默认假设的空间



# 假设验证

---

## ■ 假设检验步骤

### 2. 计算检验统计量

- Z评分

用户数60

平均花费3900

消费标准差1200

$$Z = (3900 - 4100) / (1200 / \sqrt{60}) = -1.29$$

- 将数据进行标准化处理

### 3. 评估假设所用的临界值

- 临界值的2个因素: 假设类型 (双尾检验)

显著性水平



对应犯某个错误的概率



# 假设验证

## 3.评估假设所用的临界值

### ➤ 显著性水平

II类错误：抽样问题

I类问题（需关注）：什么错误范围可以接受

结论	真实情况	
	4100	≠4100
拒绝默认假设	I类错误★	结论正确
接受默认假设	结论正确	II类错误

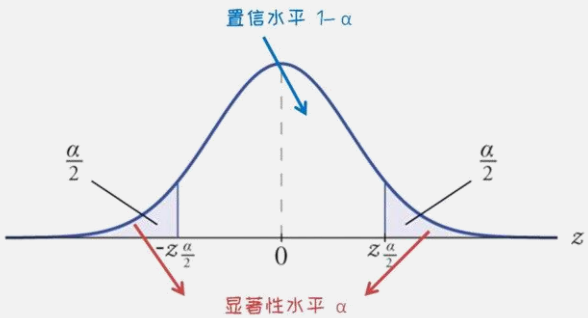
临界值是考试的及格分数，显著水平是控制多少学生及格

- 显著水平越低，考试难度越大，原假设难被否定
- 显著水平越高，考试难度越低，原假设容易被否定

显著水平定义区间通常为0.01~0.1

### ➤ 置信水平与显著性水平

- 置信水平代表了数据落在估计区间的正确率
- 显著性水平代表了数据落在估计区间的错误率
- 通常我们选择95%作为置信水平，也就是5%的显著性水平
- 对应的临界值 $z$ 为 $\pm 1.96$



置信水平	显著性水平	$\alpha/2$	临界值 $z$
95%	5%	2.5%	$\pm 1.96$
99%	1%	0.5%	$\pm 2.58$



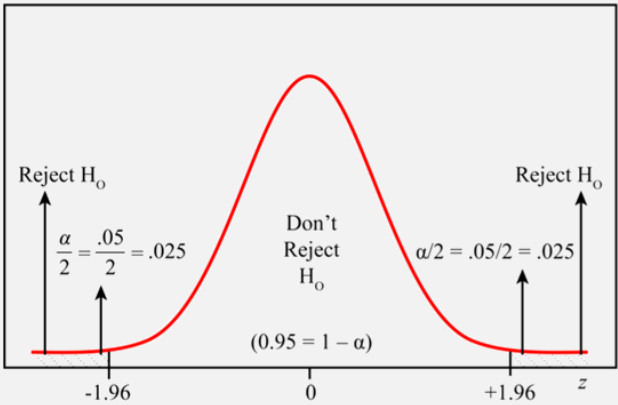


# 假设验证

## 3.评估假设所用的临界值

➤ 临界值

- 默认假设 = 4100
- 显著性水平0.05
- 双尾检验，概率水平均分， $0.05/2=0.025$
- 查找表，通过0.025，找到第一列±1.9，第一行0.06，临界值为±1.96



如果用图来表示，我们要观察的就是Z评分和±1.96的关系

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170

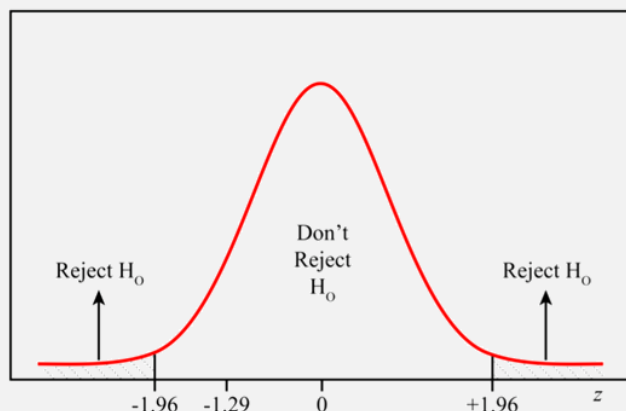


# 假设验证

## 4.做出决策判断

### ➤ 结论

- $Z=1.29$
- 临界值 $\pm 1.96$
- 结论：不拒绝，也就是用户的平均花费在4100元



现在我们有Z评分1.29，也有了临界值 $\pm 1.96$ ，我们的比较就一目了然。我们则没有足够的数据可以拒绝原假设。

### ➤ 总结问题以及解决的步骤

1. 根据我们要验证的业务问题“用户消费的均值是否就在4100元”，建立我们的默认假设
2. 根据样本数据计算出Z评分
3. 设定我们的显著水平以及根据双尾/单尾检验，得到我们的临界值
4. 比较Z评分和临界值，并给出我们的决策判断。



# 检验组间差异

在实际业务中，比较两个不同样本之间的问题

- 给用户推送的微信文章中，采用了两种不同风格的主题文案，然后收集在六组用户中的阅读量
- 这两个文案之间哪个更好？

➤ 方差分析：用于两个及两个以上样本差别的显著性检验

例子中关键的因素：文案的区别

分析差别：做出必要假设

假设：数据是服从正态分布，而且不同组之间不受到其他因素的影响

做下一步的分析

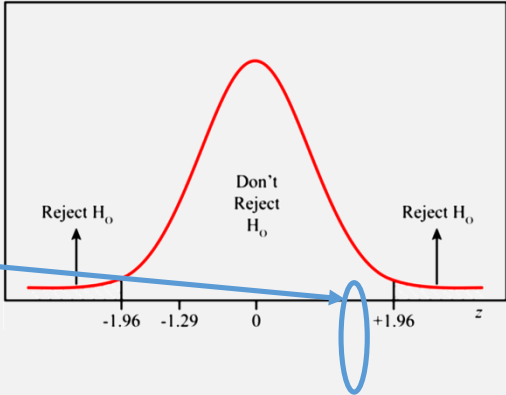


文案A	2000	1600	1900	2200	....	1800
文案B	1900	2000	1700	1700	...	1900

# 检验组间差异

➤ 方差分析具体流程（与假设验证极为相似）：

- 先做出两组之间无差异的假设，并服从正态分布
- 计算样本之间平均值的差异，构建我们需要检验的统计量 (Z)为1.03
- 选择合适的显著水平 (0.05) 和临界值 ( ±1.96 )
- 比较两者之间的大小后，去判断是否接受默认假设



$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Step1:  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{75} + \frac{(150)^2}{70}} = 29.2363$

Step2:  $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(1,480 - 1,450) - 0}{29.2363} = 1.03$  之前做出无差别的假设，为0

	发送	平均阅读量	标准差
文案A	75	1480	200
文案B	70	1450	150

结论：A和B之间在阅读量方面并没有足够明显的差异



# 检验组间差异

---

## ■ 方差分析核心

将我们所看到的数据差别

分解成为不同组之间的差异和各个组之间内部所产生的差异

如果组之间的差异大到某一个特定的比例

⇒ 可以认为是我们所关注的变量产生了足够大的影响





# 数据分析

## 第22课：用聚类分析实现客群分层

# 什么是聚类分析

---

- 定义：聚类分析是把**相似**的分析对象**根据各自特征**分成不同的组别的统计方法。
- 最常见的聚类分析应用场景，是**客户分群(segmentation)**，并由此衍生出对客户画像工作。



# 什么是客户分群

---

## ➤ 客户分群的目的:

利用顾客特征属性将顾客总体分成若干顾客群组

使得**组内**顾客特征相似

同时**不同组**的顾客之间的特征差异较为明显

## ➤ 客户分群的数据维度:

消费者行为习惯数据

消费者对产品的态度

消费者自身的人口统计学特征

顾客们消费行为的度量如RFM(时长/频率/金额)等数据





# K均值聚类分析方法

---

## ➤ 核心

将所有的观测之间划分到K个群体

使得群体和群体之间的距离尽量大

同时群体内部的观测之间的“距离和”最小

## ➤ 是一种快速聚类法

采用该方法得到的结果比较易懂

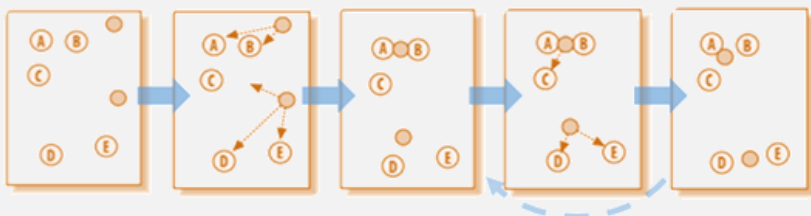
对计算机的性能要求不高

因此应用也比较广泛



# K均值聚类分析方法

## ■ 具体工作流程：



设定分组的群数



然后随机制定这个群组的中心



将离中心最近的个体归到相应的群组



重新计算群组的中心点



用新的中心对个体进行再归组

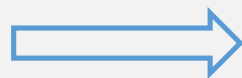


# K均值聚类分析方法

---

## ■ 计算数据之间距离——聚类分析的核心

最典型的计算距离的方式



$$\sum_{i=1}^n (x_i - \bar{x})^2$$

距离越大相似度越低，距离越小相似度越高



# 聚类分析输出案例

## 案例背景

- **企业：**旅游企业的用户数据
- **目的**
  - 希望能通过聚类分析**实现客户分群**
  - 实现对不同客户的**差异化策略**
  - 制定**针对性的旅游产品**
- **数据【表格】**
  - 自身数据
  - 第三方公司合作

数据维度	变量
基本信息	性别
	年龄
	婚姻状态
	教育水平
经济状况	是否买房
	家庭年收入估计
	总资产估计
兴趣爱好	电子游戏
	电影娱乐
	美妆时尚
	体育健身
历史消费	最近一次出行时间
	最近一次出行类型
	是否有带小孩出行

# 聚类分析输出案例

数据维度	变量	分组
基本信息	性别	1 女
	年龄	3
	婚姻状态	2 已婚
	教育水平	4
经济状况	是否买房	3
	家庭年收入估计	2 收入高
	总资产估计	1 资产多
兴趣爱好	电子游戏	4
	电影娱乐	3
	美妆时尚	1 喜欢
	体育健身	1 喜欢
历史消费	最近一次出行时间	2
	最近一次出行类型	3 一月内
	是否有带小孩出行	2

## 表格说明：

### ➤ 群组1的客户

爱好美妆和体育的高收入女性，是典型的“白富美”用户群体。

### ➤ 群组2的客户

几个变量结果是这群用户结婚并且有小孩居多，受教育程度也比较高，这群人比较富裕也热爱享受高品质生活，是高品质境外游的重要潜在客户，这就可以让相对应的业务部门制定出新的营销计划。

# 聚类分析落地效果

---

## 聚类分析是非常注重落地效果的数据分析方法

每次我们对用户或者产品或者社区进行聚类，都应该问下面这几个问题：

- 聚类之后的用户分群是否有明显的特征？
- 聚类之后用户分群是否有足够数量的用户？

如果你的用户分群结果中，有一个人群只有5个客户，那这种分群是毫无实际意义的，因为我们的营销活动不能指望五个用户的消费来实现大的盈利。

- 这些分群是否能够被触达？

**分群结果必须是可操作的。**如果我们分群中出现了很多无法触达的客群，比如没有手机号，微信或者邮件等任何一个联系方式，那么即便再有价值的分群，也是对企业来说没有可操作空间的。





# 数据分析

## 第23课：用回归分析预测销售额变化

# 回归分析的应用场景

## 媒体广告投放效果研究：

### ■ 公司可以投入的营销渠道



#### ➤ 传统大众媒体

电视、广播、户外



#### ➤ 直销媒体

电子邮件、短信、电话



#### ➤ 数字媒体

App、微信、社交应用

- 媒体的形式越来越多样化，影响受众行为的方式也日益增多





# 回归分析的应用场景

---

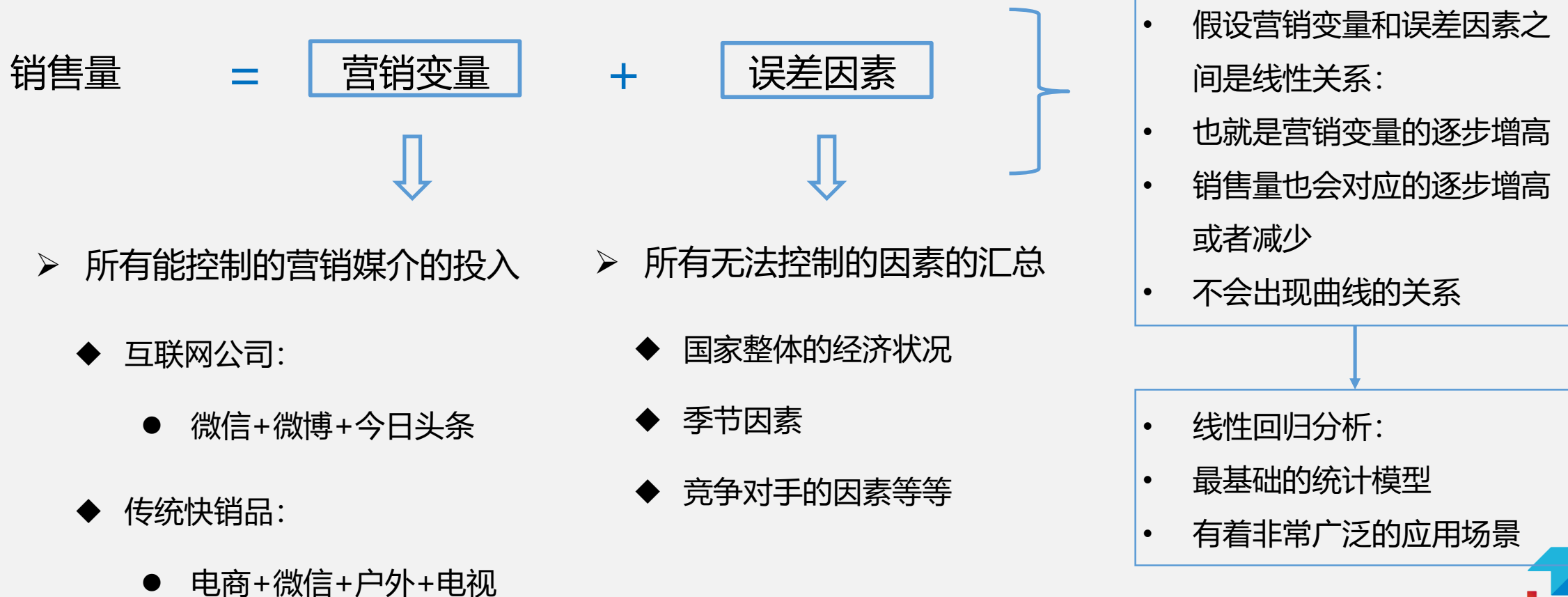
## ■ 业务分析希望能帮助回答的问题：

- 各个媒体是如何相互影响并促进销售的？
- 如何调整媒体组合从而最大化每一份支出的收益？
- 如果我们同时进行两个广告营销活动，如何判断其中一个是否比另一个更有效？



# 回归分析的着眼点

## ■ 从回归模型的角度，我们的分析着眼点可以简化为下列关系：



# 线性回归分析

---

## ■ 它是最基础的统计模型，有着非常广泛的应用场景：

- 保险场景中，用户的保费和赔付金额
- 娱乐场景中，用户的出行次数和度假市场
- 电商环境中，用户网页停留时间以及购物车的商品数量等

## • 数字型的变量，都可以作为回归模型中的目标



# 数据的准备

## ■ 分析目标：以最少的媒体投入获得最高的销售额增长

### ➤ 准备的数据：

Y：销售额

X：电视广告投入

社交媒体投入

电话直销投入

短信推送投入

线下媒体投入

➤ 回归分析中，Y和X的观测值应该在同一个时间单位

如果各种广告媒体都是以**月度**来记录投入的话，对应的销售额也  
应该是在**月度**的记录上

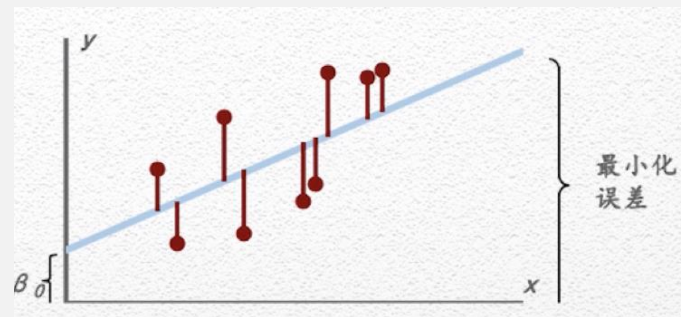
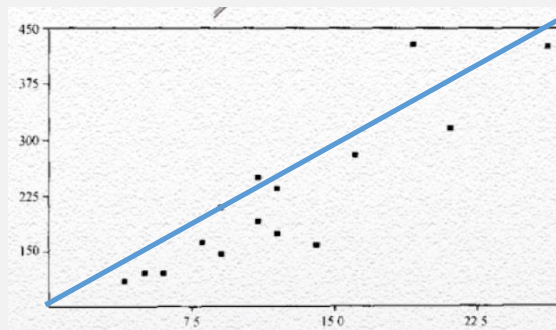
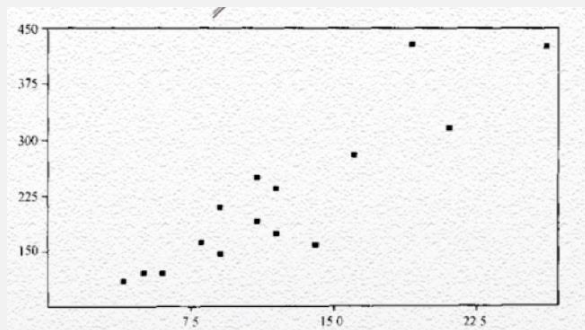
➤ 从分析的精确度上，会希望观测的时间点越多越好

比如收集最近一年的数据，广告和销售能达到**每周**或者**每日**的数据



# 工作流程

## ■ 分析目标：以最少的媒体投入获得最高的销售额增长



### ➤ Step1:对数据进行散点图制作

- x轴-电视广告的销售数据
- y轴-销售额的数据
- 得到上面若干个数据的散点图
- 随着解释变量的提高，目标变量也在逐渐提升

### ➤ Step2:引入回归线

- 回归线：对两个变量相关关系的总结
- 对于那些我们没有观测数据的地方
- 如果电视广告取到了x轴上某个数字，对应的销售额也应该是在这条线所对应的y轴的位置

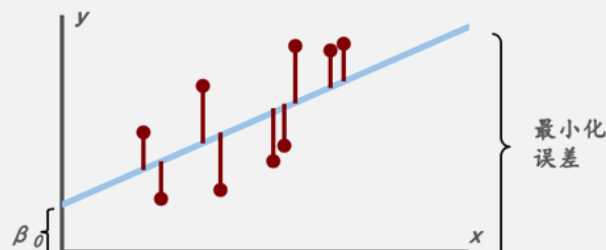
### ➤ Step3:评估回归线

- 每个点所描绘的回归线之间的距离，即是回归分析和实际值之间的差别、
- 差别越小，就代表了我们的回归分析的结果越准确



# 线性回归分析

## ■ 分析目标：以最少的媒体投入获得最高的销售额增长



➤ 准备的数据： Y：销售额 X：电视广告投入

$$Y = \beta_0 + \beta_1 X_1$$

$\beta_0$  截距项：可以理解为如果不投入任何电视广告，预计销售额会有多少

$\beta_1$  斜率：可以理解为如果多投入1单位电视广告，预计销售额会增加多少，反映了电视广告对销售额的影响程度

$$Y = \beta_0 + \beta_1 X_1 + \boxed{\varepsilon} \Rightarrow \text{加入随机误差项 } \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_p X_p + \varepsilon$$

如果我们想要同时考虑多种营销渠道对销售额的影响的话，我们可以在上面一元线性回归方程中加入更多的解释变量



# 线性回归分析

---

## ■ Step4: 扩展状态

- 将一个变量 $x$ 扩展到所有变量 $y$ 的影响
- 统计软件会给每个变量提供相对应的 $p$ 值=验证统计显著性的 $p$ 值
  - 相对应的 $p$ 值：评估回归分析中的各个解释变量，是否有对目标变量有显著影响；如果不显著，那么这些解释变量就不该保留在模型中



# 线性回归分析

---

## ■ 模型能够解释真实生活中的环境，就来自于两个因素：

- 一个是我们所能观察到的数据所造成的影响，本例中就是各个媒体投入的数据
- 另一个则是我们那些无法观测无法控制的因素所造成的影响，这些代表了市场和经济方面我们无法控制的因素

- 数字型的变量，都可以作为回归模型中的目标





# 线性回归分析

---

## ■ 回归分析是最常见的分析方法，背后又很复杂的统计假设：

- 比如最重要的一个假设：随机误差项是一个平均值为0的随机变量而且服从正态分布
- 在完成回归分析的过程中，要进行对应的验证



# 回归分析的结果

---

## ■ 回归分析的结果通常会以y和x之间的方程来描述

- 线性回归：假设**解释变量**和**因变量**之间是线性关系
  - 在现实中，销售收入不会随着广告的投入而直线上升
  - 使用回归模型的结果：最主要的是观察**各个因素系数**的大小，横向比较它们**对目标变量的关系**



# 总结

---

- **回归分析的目的：**

从相关关系到因果关系，进而能实现预测

- **回归分析的结果：**

着重于不同X与Y的影响对比，而非依赖于线性关系对未来做出非常明确的预测

线性回归主要用于**统计推断**





# 数据分析

## 第24课：用分类分析预测消费者行为

# 什么是分类分析？

## ■ 定义：

根据现有数据中对用户或者产品等的类别特征，抽象归纳成为模型，并能为新的用户或者产品等进行类别预测的过程

## ■ 三种分析方法对比：

### 分类分析 vs 回归分析

- 不同点：
  - 分类分析针对**类别型变量**
  - 回归分析针对**数字型变量**
- 相同点：
  - 都是通过X值预测Y值

### 分类分析 vs 聚类分析

- 不同点：
  - 聚类分析的目的：将一系列点分成若干类型，**类别是未知的**
  - 分类分析的目的：为了确定一个点的类别，**类别是已知的**
- 相同点：对于想要分析的目标点，都在数据集中寻找离它最近的点



# 分类分析的应用

---

## ■ 消费者行为预测：

- 注册阶段：顾客是否会参与市场活动
- 留存阶段：顾客是否会购买
- 流失阶段：顾客是否会流失

## ■ 常见应用场景：

- 判断是否为垃圾邮件
- 判断在线交易是否存在风险
- 判断用户信用卡还款



# 如何实现分类分析

---

## ■ 逻辑回归模型：结果所生成的打分方法，有着很强的业务指导性：

- 基于模型的预测值对数据进行打分
- 打分的区间在0和1之间
- 分数越高，代表目标变量变为1的可能性越高





# 数据分析

## 第25课：组间差异评估营销效果



# 组间差异分析

➤ 定义：组间差异分析也被称为AB测试

- 将某个产品/方案/设计的两个不同版本随机展示给类似的用户群体
- 以各组之间的效果差异来评估选择更好的那个

➤ 应用场景

- 分类分析，已经找到更好的客户  
将这些客户放到“营销组”中：收到广告、优惠券等
- 设定“对照组”

营销活动方法	组	数量	响应者	响应率
微信主题推送	营销组	219,501	14,721	6.7%
	对照组	41,712	2,336	5.6%

在提供同等营销刺激的情况下，好客户是否会得到更好的转化效果？

# 组间差异分析

➤ 使用的数据会是什么样

以某个快销企业提供的优惠券测试效果为例

**目的：**该公司为了促进微信端优惠券的使用

**动作：**将60万用户随机分成三组

统计为发放的三个不同的优惠券的使用率

为以后定期的优惠券发放进行测试

**数据对比：**表面上 200-20和100-10的优惠券要好于50-5的优惠券

**慎重的分析** 如果要从前面两个选一个去推广给全体客户，  
到底改选那个版本？

有时会拿到更底层的数据

优惠券200-20	8.1%
优惠券100-10	7.9%
优惠券50-5	7.2%

优惠券	用户id	兑换状态
优惠券200-20	1001	1
优惠券100-10	1002	0
优惠券50-5	1003	1

# 组间差异分析

■ 组间差异分析：

- 使用假设检验，检验两个数据之间是否有明显的均值差异
- 计算z评分，汇总各个组的均值及历史数据中 所推断的各种标准差数据

■ 具体的工作流程：

用假设验证的方式

- **先做个假设：**200-20和100-10的优惠券在用户的接受的上没有差别
- **面对这个假设，实际要回答的问题：**  
接受度相同情况下，样本之间出现0.2%的概率有多大？

从200-20和100-10的两种优惠券方式，  
分别抽取18组用户，进行组间差异分析

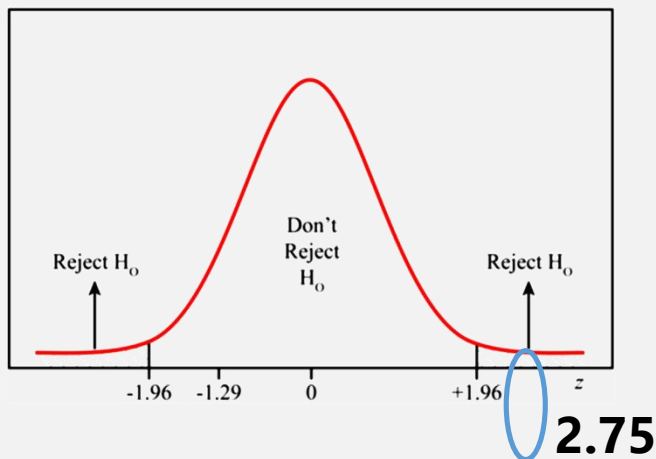
小组	200-20组	100-10组
1	0.08272682	0.0777136
2	0.0828031	0.0809198
3	0.0781537	0.0807196
4	0.08114053	0.0805868
5	0.08118523	0.0767228
6	0.08268215	0.0777152
7	0.07807296	0.081557
8	0.08284772	0.0815004
9	0.08383554	0.0756489
10	0.07746301	0.0804105
11	0.08375365	0.0825114
12	0.08058379	0.076476
13	0.07828457	0.0758352
14	0.07667982	0.0806545
15	0.08391758	0.0793583
16	0.0826379	0.0776834
17	0.08385873	0.0760742
18	0.08162487	0.0807311

# 组间差异分析

## ■ 具体的工作流程

### 假设检验步骤：

- 设定默认假设：优惠券使用情况没有显著差异
- 计算得到Z评分：2.75
- 2.75超过了0.05显著水平和双尾检验的得到的1.96临界值
- 拒绝原假设
- 在业务上，我们认为200-20的优惠券更有吸引力



### Z评分计算过程：

维度	200-20组(1)	100-10组(2)
方差 $\sigma$	0.0025	0.0023
均值 $\bar{x}$	0.0812	0.0790
样本量 $n$	18	18

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{0.0025^2}{18} + \frac{0.0023^2}{18}} = 0.000801$$
$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(0.0812 - 0.0790) - 0}{0.000801} = 2.75$$



# 组间差异分析

---

## ➤ 应用价值

- **组间差异分析有助于检验前面所做的客户分群，客群筛选结论**

最终得到的结论都围绕着“是否有显著差异”这个核心问题

- **分析目标**

可能包括广告，优惠券，微信推送，网页版本等任何可以调整测试的目标



对于用尽可能少的营销投入产生最大化的效果，有重要的指导意义

如：在互联网环境下，由于用户群体的巨大

如果能在AB测试中找到千分之一的响应率的提升

都很有可能对业务产生比较大的优化作用



# 组间差异分析

## ➤ 应用价值

做好组件测试的三个步骤：

- 一、明确我们要测试的目标
- 二、尽可能避免任何其他因素的干扰
- 三、将测试客户随机的分散到两组

**挑战：在传统营销环境中，组件分析遇到了很多挑战**

如：宝洁想测试同一个洗发水的两个包装哪个能带来更好的销量



**挑选目标对比测试销售效果** 沃尔玛或者大润发的几个商店提供新包装产品

- 其他几个继续使用老包装



**干扰因素非常多**

- 不同商店之间该产品销量原本就有差别
- 测试期间各商店可能有自己的促销活动





# 数据分析

## 总结

# 数据分析所需要的统计学基础概念

## ➤ 描述性统计：帮助理解数据的概况

- 通过对集中度和离散度的衡量
- 能够用精简的信息实现对数据的描述

## ➤ 概率和概率分布

- 帮助定义事件发生的可能性
- 做出预测的基础

## ➤ 通常所做的预测可以分为两个方面

- 根据手上的数据样本，来预测总体数据的指标
- 先对总体做出某种假设，再用数据样本进行验证

这些统计理论和验证方法为今天所流行的数据分析方法提供了理论基础





# 数据分析所需要的统计学基础概念

---

## ➤ 四个分析方法：

- 用于对用户进行分组并给出画像的聚类分析，
- 用于预测销售状况作出营销预算规划的回归分析，
- 用于筛选优质客群实现营销投入最大化的分类分析，
- 用于检验不同广告，营销，推广活动效果之间差异的组间差异分析

