【总览】

1.数据来源 ---- 抽样和抽样来源

2.数据类型

3.如何描述数据 ---- 使用表格、图形来汇总数据

使用数值来描述数据

4.数据分析 ----概率的计算，分配，性质，定理

概率分布（离散型，连续型）

☆☆☆☆☆☆☆☆☆☆☆ 数据来源 ☆☆☆☆☆☆☆☆☆☆☆

数据的来源：

1.现有来源 ----- 公司内部数据，专门搜集保存的机构，行业协会盈利机构，互联网，政府机构

2.观测性研究 ----- 选择样本，记录统计数据，比如调查与民意调查，无法控制变量

3.实验 ----- 选定特殊变量，保持其他变量不变，以获得该变量的影响，这个是可控的变量

注意：必须清楚获得数据所需要的时间与成本，尽量使用短期现有数据，不然需要考虑时效以及付出成本，结果的可靠性

数据采集有误差，需要甄别

========== 抽样和抽样来源 ==========

这讲的是通过样本的均值，标准差，样本比率参数，得到总体的预估值，同时给出预估值正负 XX 范围的概率

不是讲怎么抽样准确率高，不要被骗了！

抽样总体：抽取样本的总体

抽样框：用于抽取样本的个体清单

参数：总体的数字特征，例如平均值，标准差

1.抽样

(1) 有限总体 ----- 无放回抽样：忽略已被选取过的样本

有放回抽样：对已经出现过的样本仍选入

简单随机样本：从容量为 N 的有限总体中抽取一个容量为 n 的样本，并且容量为 n 的每一个可能的样本都是等概率的被抽出

通常是编号 + 随机数的方式

(2) 无限总体 ----- 建议抽取一个所谓的随机样本，但是有如下：

随机样本的条件：1.抽取的每一个个体来自同一个总体

2.每个个体的抽取是独立的

条件 1 简单，条件 2 主要是防止偏差，例如错误的选择顾客都是 60 岁以上的等

(3) 样本比率 $\bar{p}$ 是总体比率 p 的点估计，计算公式为： $\bar{p} = x / n$ x-样本中具有特征的数量，n-样本容量

意义是考试合格率，总体是 60%，可能样本中只有 57%

2.抽样分布

假设我们对总体进行了一次抽样，确定了样本的平均值 $\bar{x}$，如果再来 500 次抽样，再得到 500 个平均值 $\bar{x}$，那么

平均值 $\bar{x}$ 的取值也有各种可能性，我们称 $\bar{x}$ 的概率分布为 $\bar{x}$ 的抽样分布，即

$\bar{x}$ 的抽样分布是样本均值 $\bar{x}$ 的所有可能值的概率分布

$\bar{x}$ 的数学期望： $E(\bar{x}) = \mu$ μ 为总体均值

$\bar{x}$ 的标准差：

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

有限总体 无限总体（n/N<0.05 可以当成无限）

σ - 总体标准差 n - 样本容量 N - 样本总量

作用是反推总体期望与标准差（不过这是 $\bar{x}$ 的抽样分布）

分布形式：有服从正态分布与不服从正态分布 2 种形式，但是

中心极限定理：从总体中抽取容量为 n 的简单随机样本，当样本容量很大时，样本均值 $\bar{x}$ 的抽样分布近似服从正态分布

一般 n 大于 30 即可，若总体是严重偏态或是异常点，则要 50

应用：假设抽查的样本容量是 30，均值是 $\bar{x}$，老板想要知道样本均值 $\bar{x}$ 在总体均值 ±500 的概率有多大

因为容量>30，可以近似认为正态分布，那么求（ $\bar{x}$ -500， $\bar{x}$ +500）的概率，知道均值 $\bar{x}$，知道标准差，就能算，详见正态分布概率算法

如果觉得概率太低，可以增加样本容量，那么均值 $\bar{x}$ 的标准差在减小，算出来的概率更具有可靠性

$\bar{p}$ 的数学期望： $E(\bar{p}) = p$

$\bar{p}$ 的标准差：

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

有限总体　　无限总体（n/N<0.05 可以当成无限）

p-总体比率，N-样本总量，n-样本容量

分布形式：当 np>=5 并且 n(1-p)>=5 时， $\bar{p}$ 的抽样分布可以用正态分布

应用：假设样本合格学生占 60%，那么老板想知道全部考生合格率在 55%-65%之间的概率是多少

算法与上面的一样，同样样本容量越大越好

☆☆☆☆☆☆☆☆☆☆　数据类型　☆☆☆☆☆☆☆☆☆☆

分类型数据 ----- 归属于某一类别的数据，里面是分类变量，可以用名义尺度，顺序尺度

一般统计分析方法有限，需要计算对比每一类别中的观测值的数目

数量型数据 ----- 用于表示数量大小多少的数值，里面是数量变量，可以用间隔尺度，比率尺度

截面数据 ----- 在相同或近似相同的同一时间点上搜集的数据

时间序列数据 ----- 是在几个时期内搜集的数据

☆☆☆☆☆☆☆☆☆☆　如何描述数据　☆☆☆☆☆☆☆☆☆☆

========== 使用表格、图形来汇总数据 ==========

【总览】

汇总分类变量的数据　　汇总数量变量的数据

表格法汇总两个变量的数据

图形显示法汇总两个变量的数据

最终的数据可视化

【汇总分类变量的数据】

1.频数分布 ----- 表格法，表示在几个互不重叠的组别中，每一组项目的个数（即频数）

展示：　类别　频数　相对频数分布：0.38　百分数频数分布：38%

2.条形图 ----- 图形法，用于描绘分类型数据的频数分布，相对或百分数频数分布

横轴：类别

纵轴：频数或者相对，百分比频数

3.饼形图 ----- 图形法，用于描绘分类型数据的相对频数和百分数频数

圆分成若干个扇形，每个扇形与每组数据的相对频数对应

注意：一般来说条形图的展示效果比饼形图好

条形图，饼形图还可以考虑颜色，阴影，文本，三维透视图的信息增加效果

可以把频数较小（5%或更少）的类别合并到“其他”的综合组去

【汇总数量变量的数据】

1.频数分布-----类似上面的，但是必须小心定义这个互不重叠的组别

组数：5-20，根据数据项来选择，比如数据量是 20，那么我们可以设置 5 组

组宽：一般每组的宽度相同

近似组宽 =（数据最大值 - 数据最小值）/ 组数　一般小数往上取整，4.2 就取 5

组限：（最小值，最小值 + 组宽），然后依次往上计算，例如 10-14，15-19

组中值：上组限与下组限的中间值，例如 10-14 那么就是 12，有时可能有用

累计频数分布：组别成了小于等于 14，小于等于 19 这样的，解决了“20 天完成的达标率是多少”这样的问题，例如<=19

2.直方图 ----- 相比条形图，直方图柱子与柱子之间是相互接触的，当然也可以分开

横轴：所关心的变量，例如 10-14，15-19 这样的组

纵轴：频数或者相对，百分比频数

3.打点图 ----- 画一条数轴，上面根据数据打点，这个折线图能替代吧

4.茎叶图 ----- 画一条竖线，左边是 5,6,7,8,9，右边是对应分数的个位数，有点像折线图，茎|叶，叶单位可以=10,100

优点是便于绘制，其次能提供更精确的信息

注意：

设置组限要考虑数据的精度

可以根据情况设置开口组 --- “35 或 35 天以上”

【汇总 2 个变量的数据 --- 表格法】

1.交叉分组表 ----- 表的横与列分别代表 2 个变量，中间的数据则是符合这 2 个变量的样本的频数（依这点来区别普通的频数分布）

2 个变量分别在左边和上方哦，别放在下方，那里是统计总数的地方，数据若是 0 可以省略不写

任何行，列末尾都可以加频数总计，相对频数总计，百分数频数总计（也可以在总计，或者干脆在数据后面加百分比）

通过行列中数据，可以看出趋势，例如随着 X 轴的增加，Y 轴 XXXXXX

一个谬论：当我们通过 2 张交叉分组表，组合成一张交叉分组表时，必定会省略一个因素，得到的结论可能会与分开看相反

例如分组选举的谬论，因为多数服少数的分组原则，即使总支持人数低，也能通过适当分组，最终获选

【汇总 2 个变量的数据 --- 图形法】

1.散点图 ----- 必须是 2 个“数量变量”，然后一横轴一纵轴，开始打点，通过计算协方差，相关系数，可以知道斜率

趋势线 ----- 依据打点，试着画一条线，一般是正相关，负相关，无关系

2.复合条形图 ----- 一般是一个“数量变量”，一个“分类变量”

数量变量为横轴，频数为纵轴，3 个分类变量就画 3 个柱子，可以由颜色区分并备注分类颜色

从某一分类的趋势看，分析另一个分类的趋势，是否有关系

结构条形图 ----- 一般以数量变量为横轴，频数百分比为纵轴，每个数量变量就 1 个 100%的柱子，内部用不同颜色表示不同分类变量占的百分比

通过分类变量占的百分比的趋势，可以归纳规律

也可以用频数来替代纵轴的频数百分比，这样柱子就有高低

【最终的数据可视化】

用于展示分类型数据的频数分布和相对频数分布	分类型数据	条形图
用于展示分类型数据的相对频数分布和百分数频数分布	分类型数据	饼形图
用于展示数值型数据在整个数据范围内的分布	数值型数据	打点图
用于展示数值型数据在一个区间组集合上的分布	数值型数据	直方图
用于展示数值型数据的等级顺序和分布形态	数值型数据	茎叶显示
用于两个变量的比较	两个变量的比较	复合条形图
用于比较两个分类变量的相对频数和百分数频数	两个分类变量	结构条形图
用于展示两个数量变量的相关关系（记得试试画趋势线）	两个数量变量	散点图
多个箱形图组合汇总数据		箱形图
多张图放在一张大图里，一目了然，减少屏幕滚动，方便查看		数据仪表盘

========== 使用数值来描述数据 ==========

【总览】

对位置的度量：　　平均数，中位数，众数，加权平均数，几何平均数，百分位数（四分位数）

对变异，离散程度的度量：　　极差，四分位数间距，方差，标准差，标准差系数

数据的分布形态度量：　　偏左或偏右，利用经验法则或是切比雪夫定理来为数据分布提供更多信息，以及识别出异常值

汇总的度量--箱形图：　　同时提供数据分布位置，变异程度和形态

衡量 2 个变量间的关系：　　协方差与相关系数，就是看线性关系的大小程度

【位置的度量】

1.平均数 ----- 提供了数据中心位置的度量

$\bar{x}$：样本平均数，其中样本总数量用 n

μ：总体平均数，其中总数量用 N

我们把 x_1 代表变量 x 的第一个观测值，加起来除以总数量 n 就得出了平均数

特殊：

调整平均数 ----- 剔除 5%的最大，最小值，将剩下值计算平均数，那么结果就是 5%调整平均数

加权平均数 ----- 每个观测值所占的权重不同，那么在计算平均数的时候就要算上权重

例如购买多批次的相同物品，每次的价格乘以数量，同一加起来除以总数量，与普通平均数多了几步

2.中位数 ----- 对变量中心位置的另一种度量

奇数个观测值，就是中间的数值，偶数个观测值，就是中间 2 个数值的平均值

在平均数受到异常大异常小的数值影响时，中位数更适合。当然，有时可以去除极端值

3.几何平均数 ----- 对于乘法过程，诸如增长率的应用，几何平均数是合适的位置度量，0.94，1.02 都是增长因子

公式为：n 个增长因子乘积的 n 次方根 - 1，

如时期 1，2，3 回报率分别为-6%，2%，-3%，那么为 $(\sqrt[3]{(0.94*1.02*0.97)} - 1) * 100\%$

例如某个基金的年增长率，相加除以总数就错了，应该相乘(1 + 每年的增长率)取 n 次方根，最后值-1，乘以 100%

计算第 3 年的，就是 1*(几何平均数)的 3 次方，或者 1*0.94*1.02*0.97 的这样算

适用于财务，投资，银行业各种几个连续时期的平均变化率，还有出生率，死亡率，年，月变化率

这跟年数根号（最后一年/最初一年）一样

4.众数 ----- 出现次数最多的数据

正好有 2 个众数，就是双众数，若是 3 个，则是多众数，但是太多对描述数据不起多大作用

5.百分位数 ----- 第 p 百分位数将数据分割成了 2 个部分，提供了数据如何散步在最大，最小值的区间上的信息

第 p 百分位数 = p * (n + 1) / 100

将数据从一定规则排列（从小到大），若是 10 个数据，那么第 80 百分位是 8.8，具体数值为：第 8 + （第 9 - 第 8）* 0.8

例如 630 分对应第 82 百分位数，则代表 82%的学生分数比 630 分低，18%的比这个分数高

四分位数 ----- Q_1 = 第一四分位数，或第 25 百分位数

Q_2 = 第二四分位数，或第 50 百分位数（也是中位数）

Q_3 = 第三四分位数，或第 75 百分位数

作用是可以形象地将数据展示为四部分：

25%的数据 | 25%的数据 | 25%的数据 | 25%的数据

Q_1 Q_2 Q_3

其他常用的是五分位数，十分位数

【变异程度的度量】（就是离散程度）

1.极差 ----- 最简单的变异程度的测量（很少被单独使用，极易受到异常值的影响）

极差 = 最大值 - 最小值

2.四分位数间距 ----- 值为：第三四分位数 - 第一四分位数，即 $Q_3 - Q_1 = IQR$

反映的是中间 50%数据的极差，能较好的克服异常值的影响

3.方差 ----- 用所有数据对变异程度所做的一种度量，依赖于每个观察值与平均值之间的差异

方差 = 每个（观测值 - 平均值）的平方全部相加的结果 / 总数

样本方差：$s^2 = \sum(x_i - \bar{x})^2 / (n-1)$ # 总之减 1 是对总体方差的无偏估计

总体方差：$\sigma^2 = \sum(x_i - \bar{x})^2 / N$

方差大的变量表示其变异程度也大

方差的单位也要平方，例如 s^2=64(人)²，直观很难理解，所以是比较多个变量变异程度的有用工具

4.标准差 ----- 方差的正平方跟，优势是与原始数据单位度量相同，容易与平均数什么的做比较

样本标准差：$s = \sqrt{(s^2)}$

总体标准差：$\sigma = \sqrt{(\sigma^2)}$

5.标准差系数 ----- 标准差相对于平均数大小的描述统计量

值：（标准差 / 平均数）* 100%，主要用于比较不同标准差和不同平均数的变量的变异程度

可能 2 组数据表面上 A 的标准差小，但是看标准差系数，可能反而是 A 较大，B 稳定

例如跑 500 米所需时间与 2000 米所需时间，虽然 500 的标准差小，但是 2000 的标准差系数小，跑 2000 米所需时间更有一致性

【分布形态、相对位置以及异常值的检测】

1.分布形态 ----- 可以用直方图形象的展示出来

正常应该是左右对称的，如果相对于最高的柱子，其右边的柱子数量多，则是右偏

计算方式复杂：偏度 = $n/(n-1)/(n-2)*\sum((x_i - \bar{x})/s)^3$，>0，则是右偏，通常平均数比中位数大

右偏反映了少数大值将平均数拉大，严重偏离时，中位数比较适合，当然也可以调整平均数

2.z-分数 ----- 确定一个数值距离平均数的相对位置大小，用 z_i 表示，z_i = -0.5 表示 x_i 比 $\bar{x}$ 小 0.5 个标准差

$z_i = (x_i - \bar{x}) / s$，$z_i$ 代表 x_i 的 z-分数，解释为 x_i 与平均数的距离是 z_i 个标准差

如果用打点图来表示样本数据，与 z-的位置，那么两张图是一样的，就是底部的单位不同

3.切比雪夫定理 ----- 知道标准差，平均数，可以算出至少有多少百分比的数据值介于 XXX-XXX 之间

定理：与平均数的距离在 z 个标准差之内的数据值所占的比例至少为（$1-1/z^2$），其中 z 必须大于 1

当 z=2，3，4 个标准差时，有如下：

1.至少 75%的数据值与平均数的距离在 z = 2 个标准差内

2.至少 89%的数据值与平均数的距离在 z = 3 个标准差内

3.至少 94%的数据值与平均数的距离在 z = 4 个标准差内

例如：100 名学生成绩平均值为 70 分，标准差为 5 分，那么

至少 75%的学生成绩在 60-80 分内，因为 60，80 距离平均值 70 为 2 个标准差

58-82 分：（82 - 70）/ 70 = 2.4 距离 2.4 个标准差

1 - (1/2.42) = 0.826，说明至少 82.6%的学生成绩在 58-82 分之间

4.经验法则 ----- 如果数据集符合正态分布，那正常来说，有：

1.大约 68.26%的数据值与平均数的距离在 1 个标准差内 -- 68%

2.大约 95.44%的数据值与平均数的距离在 2 个标准差内 -- 95%

3.大约 99.74%的数据值与平均数的距离在 3 个标准差内 -- 几乎所有

例如生产线上的罐子重量通常是正态分布的，如果平均重 16 克，标准差为 0.25 克

那么有大约 68%的罐子，重量会在 15.75-16.25 克之间

经验法则优先于切比雪夫定理，当然前提是正态分布

5.异常值的检测-----异常值指的是数据集中可能包含的一个或多个数值异常大或者异常小的观测值，这就是异常值

1.可能是被人为的不小心错误记录，那么应该剔除

2.可能是一个反常的数据值，那应该被记录保留

检测方法 1：对于正态分布，或是类似的，可以用 z-分数在[-3,3]来判断，在这个区间内正常，不在则要检查准确性

检测方法 2：通过第一，三四分位数与四分位数间距为依据判断 ----- 就是下面的五数概括法

下限 = Q_1 - 1.5 * IQR

上限 = Q_3 + 1.5 * IQR

不在该范围内的观测值就该检查一下

【五数概括法&箱形图】（用于确定一个大的数据集的几个特征）

1.五数概括法 ----- 最小值，第一四分位数（Q_1），中位数（Q_2），第三四分位数（Q_3），最大值

表明了数据集在多少的范围内，中间值为多少，50%的数据集中在 Q_1 与 Q_3 之间

2.箱形图 ----- 基于五数概括法的数据图形汇总

展示了 Q_1，Q_2(中位数)，Q_3

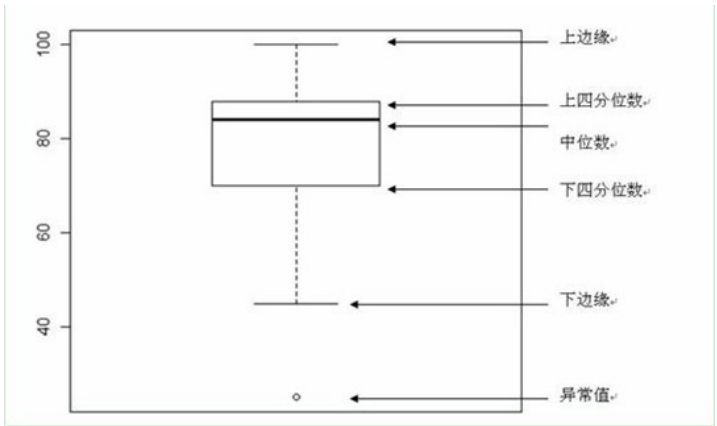
上边缘：边界内的数据集里的最大值

下边缘：边界内的数据集里的最小值

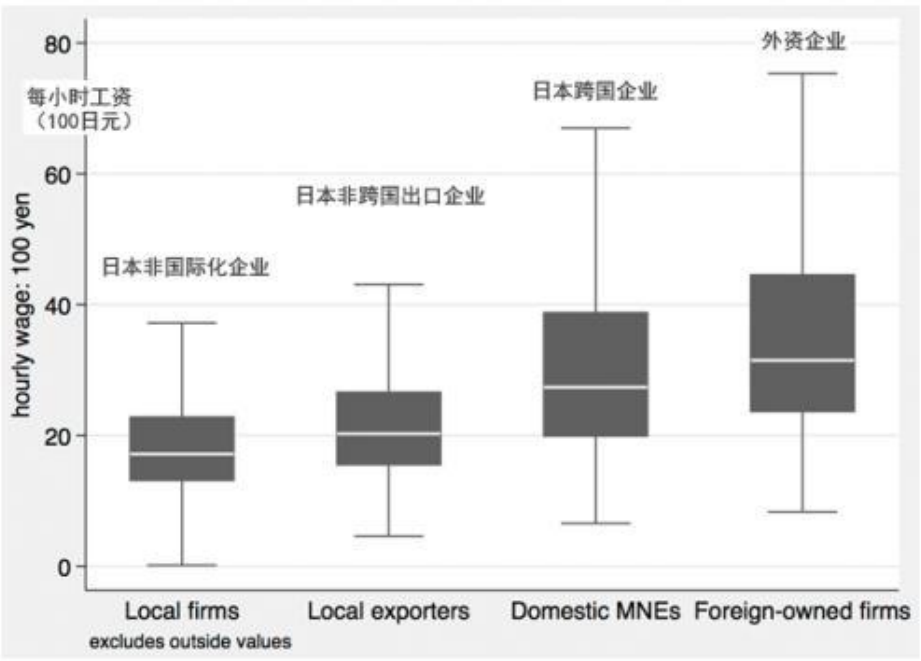
边界：| ---1.5IQR--- Q_1 ---IQR--- Q_3 ---1.5IQR--- |

异常值：不在边界内的数据

上边缘与下边缘所组成的线称为触须线



3.箱形图的比较-----多个箱形图组合可以汇总数据，箱形图高度小反应了比较稳定



【两个变量之间关系的度量】

1.协方差 ----- 两个变量之间的线性关系度量

公式：对于一个容量为 n 的样本，其观测值为$(x_1, y_1)(x_2, y_2)(x_3, y_3)(x_4, y_4)$....，那么

样本协方差： $s_{XY} = (\sum(x_i - \bar{x})(y_i - \bar{y})) / (n - 1)$

总体协方差： $\sigma_{XY} = (\sum(x_i - \bar{x})(y_i - \bar{y})) / N$　　注意总体协方差不用-1

>0：正线性关系，随着 x 的增加，y 也增加，坐下到右上的一条线

<0：负线性关系：随着 x 的增加，y 减小

近似为 0：无线性关系

2.相关系数 ------ 协方差依赖于度量单位的统一，选择米或是厘米，带来的结果完全不同，但是事实上两者的相关关系没有变化

$r_{XY} = s_{XY} / (s_X * s_Y)$　　s_{XY}：样本协方差　s_X：x 的样本标准差　s_Y：y 的样本标准差

$p_{XY} = \sigma_{XY} / (\sigma_X * \sigma_Y)$　　σ_{XY}：总体协方差　σ_X：x 的总体标准差　σ_Y：y 的总体标准差

范围是-1 到 1，越接近-1 或者 1，则线性关系越强，接近于 0 则代表没有线性关系

如果算出来相关系数等于 1，则表示存在完全的线性关系，45 度斜着出去的线

注意：1.相关系数提供了线性，但不是因果关系的度量，较高的相关系数并不意味着一个变量的变化会引起另一个变量的变化

例如饭店的质量等级和价格是正相关的，但是简单的增加价格并不会提高质量等级

2.可能存在散点图是“v”的样子，虽然相关系数趋近于 0，但是明显前半段，后半段存在着线性关系

☆☆☆☆☆☆☆☆☆☆　数据分析　☆☆☆☆☆☆☆☆☆☆

1.描述性分析：描述过去发生状况的分析技术集合

对过去的数据进行查询，描述事实

2.预测性分析：利用过去数据建立的模型来预测未来或是评估一个变量对另一个变量的影响

利用过去，分析未来

3.规范性分析：产生一个最佳行动过程的分析技术集合

得出了下一步我们的最佳做法？

========== 概率 ==========

【总览】

概率怎么算的，如何分配到某个事件

概率的基本性质

贝叶斯定理 ---- 重要，有用

【概率的算法】

1.多步骤试验计数法则 ----- 如果一个试验可以看做循序的 k 个步骤，在第 1 步有 n_1种结果，第 2 步有 n_2种结果，那么依此类推：

试验结果的总数为 $n_1 * n_2 * n_1 * ... * nk$

例子：抛硬币，抛 6 枚硬币的试验有 $2^6 = 64$ 种结果

开发需要 1 或 2 或 3 天 4 天，测试需要 1 或 2 或 3 天，那么最终可能的结果是 4*3=12 种，区间是 2 天到 7 天

工具：树形图，从 1 个点衍生 2 条线（正面，反面）到 2 个第 2 个点，再衍生 4 条线下去

2.组合计数法则 ----- 在 N 项中选取 n 项的试验，那么一共有如下种组合：

$$C_N^n = C(N, n) = \frac{N!}{n!(N-n)!}$$

其中：N! = N*(N-1)*(N-2)*...*2*1，　0! = 1

n! = n*(n-1)*(n-2)*...*2*1

例子：从 5 个零件中抽取 2 个做检查，那么有 C(5,2) = 10 种可能的结果，AB、AC、AD、AE、BC、BE、BF、CE、CF、EF

3.排列计数法则 ----- 从 N 项中选取 n 项并且考虑选取的顺序时，那么有如下种组合：

$$P_N^n = P(N, n) = \frac{N!}{(N-n)!}$$

例子：从 5 个零件中抽取 2 个做检查，并且先后顺序要考虑，那么有 P(5,2) = 20 种可能，原来的 AB 变成了 AB，BA，所以都乘以 2，为 20

【概率的分配】

大前提：所有试验的结果的概率之和必须为 1

$P(E_1) + P(E_2) + ... + P(En) = 1$　　E_i 表示第 i 种试验结果　$P(E_i)$表示这种结果的发生概率

1.古典法 ----- 在各种试验结果是等概率发生的前提下，每个试验结果发生的概率是 1/n

大前提：各种试验结果是等概率发生

2.相对频率法 ----- 在获得所有试验结果的前提下，可以计算出每种结果的发生频率

大前提：可以取得所有试验结果，意义在根据以往的数据，推算该结果发生的概率

例子：一个月下雨的日子有 10 天，那么可以说这个月下雨分配的概率是 3 分之 1

3.主观法 ----- 额，自己主观认为发生的概率，记得所有结果概率加起来要为 1

【概率的性质】

1.简单概率：

利用事件的补计算概率：$P(A) = 1 - P(A^c)$

A^c 是事件 A 的补，即所有不包含在事件 A 的样本点

加法公式 ----- 提供一种计算“事件 A，或 B，或 A,B 同时发生”的概率的方法

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$　　∪-交 ∩-交

加法公式（互斥事件）----- $P(A \cup B) = P(A) + P(B)$　　互斥事件：两个事件没有公共的样本点

2.条件概率：

注意：讨论的是 A,B 两个事件，其中事件 A 发生的概率记为 P(A)，事件 B 则是 P(B)

$P(A|B) = P(A \cap B) / P(A)$　　---代表在事件 B 已经发生的情况下事件 A 所发生的概率

$P(A \cap B) = P(B) * P(A \cap B)$　　---乘法公式，代表的是两个事件都发生的概率

如果 $P(A|B) = P(A)$　　A 与 B 相互独立，互不影响，反之则是相依的，事件 A 会受到事件 B 的影响

如果 $P(A \cap B) = P(B) * P(A)$　　A 与 B 相互独立，意义是事件 A 并不会受到事件 B 的影响

3.贝叶斯定理：

大前提：A_1,A_2 必须为互斥事件

后验概率 --- 虽然开始认为选择 A_1的概率是 65%，既然出现次品了，那么选择 A_1的概率变化成了 42%

事件	先验概率	条件概率	联合概率	后验概率
A_i	$P(A_i)$	$P(B\|A_i)$	$P(A_i \cap B)$	$P(A_i\|B)$
A_1	0.65	0.02	0.013	0.4262
A_2	0.35	0.05	0.0175	0.5738

$P(A_1|B) = P(A_1) * P(B|A_1) / (P(A_1) * P(B|A_1) + P(A_2) * P(B|A_2))$　两事件 A_1，A_2的情形

$P(A_i|B) = P(A_i) * P(B|A_i) / (P(A_1) * P(B|A_1) + P(A_2) * P(B|A_2) + ... + P(A_i) * P(B|A_i))$　　多事件

例子：A 事件：选择工厂　　B 事件：零件坏了

$P(A_i)$　　--- 发生 A_i 事件的概率　　65%的概率选择工厂 A_1

$P(B|A_i)$　　--- 在 A_i 条件下发生 B 的概率　　工厂 A_1的零件有 2%的次品率

$P(A_i \cap B)$　　--- 选择了 A_i 并且发生了 B 事件的概率　　1.3%的概率会恰好发生选择 A_1工厂偏偏出次品

$P(A_i|B)$　　--- 在确认 B 发生的情况下，选的是 A_i 的概率　　若已知出次品了，那么这个零件是 A_1工厂生产的概率是 42%

先验概率 --- 新信息来了 --- 应用贝叶斯定理 --- 后验概率（修正了之前我们认为的先验概率）

例子：5%的人用信用卡不还钱，还钱的人当中有 20%会拖一个月再还

那么如果一个人已经 1 个月不还钱了，可以通过贝叶斯定理计算最终会赖账的可能性，依此参考要不要收回信用卡

如果一个邮件包含 shopping 字段，那么他是一个广告的概率

========== 概率分布 ==========

【总览】

随机变量

离散型概率分布

连续型概率分布

【随机变量】

随机变量 ----- 对试验结果的数值描述

概率分布 ----- 是描述随机变量取不同值的概率

概率函数 ----- 给出随机变量取每种值的概率，记做 f(x)

1.离散型随机变量 ----- 可以取有限个值或者无限个值的随机变量

例如：考试有 4 门课，那么通过考试的值可能是 1,2,3,4　　-----有限个

一天中通过收费站的汽车数，可能是 0,1,2,3,....　　-----无限个

我们可以定义离散型随机变量将试验结果数值化，例如顾客性别，男 1，女 2

2.连续型随机变量 ----- 可以取某个区间或多个区间内任意值的随机变量

例如：度量时间，重量，距离，温度

两个客户到达银行的时间间隔　　x>=0

6 个月后的工作进度　　0<=x<=100

3.判断规则：任意选择随机变量的两个值，如果再这两个值之间的所有点都可能是随机变量的取值，那么这就是连续型随机变量

【离散型概率分布】

大前提：　f(x) >= 0　　---任何随机变量 x 发生的概率大于等于 0

∑f(x) = 1　　---所有变量 x 概率和等于 1

1.计算方式：古典法　----- 例如抛硬币，那么 f(1)=0.5，f(2)=0.5，f(1)代表正面，f(2)代表背面

主观法　----- 凭自己的经验给出值

经验离散分布　----- 利用相对频率法建立离散型概率分布

例如汽车的日销量，根据已知的上个月结果，可以得出

f(0) = 0.2　　1 辆车都没销售出去的概率

f(1) = 0.6　　销售了 1 辆车的概率

f(2) = 0.2　　销售了 1 辆车的概率

2.数学期望 ----- 对随机变量中心位置的度量

公式：

$E(x) = \mu = \sum(x * f(x))$　　意义是随机变量取值的加权平均，其中权数是概率

μ 是之前的平均数

例子：上个月每天可能卖出 0,1,2,3 辆车，那么通过数据计算，可以得出日销量为 1.5 辆车

方差 ----- 度量随机变量的变异或分散程度

公式：

$Var(x) = \sigma^2 = \sum((x - \mu)^2 * f(x))$　　意义是随机变量离差平方的加权算数平均，其中的权数是概率

σ^2是之前的方差

通常用标准差 $\sigma = \sqrt{(\sigma^2)}$

例如：汽车日销量为 1.118 辆，可以用于对比

3.二元概率分布 ----- 关于两个随机变量的概率分布，也被称作联合概率

例子：A 厂一天的销售汽车数量可能为 0,1,2,3 辆，B 厂一天的销售汽车数量也可能为 0,1,2,3 辆

那么 f(0,3)代表 A 厂销售 0 辆车，B 厂销售 3 辆车的联合概率

定义日销总量 s = x + y

那么 f(s=0) = f(0,0)　　A 厂销售 0 辆车并且 B 厂也销售 0 辆车的概率

f(s=1) = f(0,1) + f(1,0)

f(s=2) = f(0,2) + f(1,1) + f(2,0)

....

f(s=6) = ...

☆重要 ☆，据此，可以得出，

数学期望　E(s) = ...　代表 2 个厂期望的日销总量是多少

方差　Var(s) = ...　日销总量的变异程度

协方差　$\sigma_{XY} = (Var(x + y) - Var(x) - Var(y)) / 2$　　>0：正线性关系，随着 x 的增加，y 也增加，坐下到右上的一条线

<0：负线性关系：随着 x 的增加，y 减小

相关系数　$p_{XY} = \sigma_{XY} / (\sigma_X * \sigma_Y)$　　接近 1，正相关强，接近-1，负相关强，0 则是无关

还有，在随机变量 x 和 y 的线性组合的情形下，例如股票投 40%A，60%B

数学期望：　E(ax + by) = aE(x) + bE(y)　　就是加权平均了下

方差：　$Var(ax + by) = a^2Var(x) + b^2Var(y) + 2ab\sigma_{XY}$　　σ_{XY} 是 2 个变量的协方差

应用是 all in 股票，all in 债券，0.5 股票 0.5 债券，结果发现第三种期望收益居中，但是风险（标准差）最低

股票，债券的相关系数表明，一般股票涨了，债券就跌了

4.二项概率分布 ----- 二元概率分布的特殊情况

条件 1：试验由一系列相同的 n 个试验组成

2：每次试验仅有 2 种可能的结果，可以认为除了成功就是失败

3：每次试验成功的概率都是相同的，用 p 来表示，那么失败的概率就是 1-p，p 值不能变化

4：试验是相互独立的

那么，n 次试验中恰好有 x 次成功的试验结果的数目：　　n --- 总试验数目，p --- 成功的概率，x --- 成功的数目

试验结果的数目：

$$C_n^x = C(n, x) = \frac{n!}{x!(n-x)!}$$

特定试验结果的概率：　$p^x * (1 - p)^{(n-x)}$

故 n 次试验中恰好有 x 次成功的概率：$f(x) = n! / (x!*(n-x)!) * p^x * (1 - p)^{(n-x)}$　(就是上面 2 式的相乘)

期望：$E(x) = \mu = np$

方差：$Var(x) = \sigma^2 = np(1-p)$

注意：当 n 很大很难算的时候，如果 np>=5 且 n(1-p)>=5，可以使用正态分布来模拟该二项概率分布，详见正态分布

5.泊松概率分布 ----- 二元概率分布的特殊情况

条件 1：在任意两个相等长度的区间上，事件发生的概率相等

条件 2：事件在某一区间上是否发生与在其他区间上是否发生是独立的

例如：　一小时内到达洗车房的汽车数量

100 米长的水管有多少处发生泄漏

10 公里的路有几处需要维修

概率：$f(x) = \mu^x * e^{-u} / x!$　　x---在一个区间上发生的次数

μ---发生次数的数学期望

e---2.71828

方差与数学期望 μ 相等

如果 10 分钟出现的车辆数量的期望是 10，那么 1 分钟出现的车辆数量期望就是 1

6.超几何分布 ----- 类似二项分布，特点是：

1.各次试验不是独立的

2.各次试验中成功的概率不等　　--- 检查出一个坏的，拿出来后继续检查，那么还是坏的几率会变小

超几何分布概率：　$f(x) = C(r, x) * C(N-r, n-x) / C(N, n)$　　x 为成功的次数，n 为试验次数

f(x)为 n 次试验中 x 次成功的概率

N 为总体中元素的个数，r 为总体中具有成功标志元素的个数

均值：$E(x) = \mu = n * r / N$

方差：$Var(x) = \sigma^2 = n * (r/N) * (1-r/N) * ((N-n)/(N-1))$

例子：100 个零件，有 2 个是残次品，现在抽查小组随机选 10 个零件，恰好检查出来有坏的概率

【连续型概率分布】

我们计算的是随机变量在某个区间内取值的概率

不像离散型概率给出的是随机变量 x 取特定值的概率

1.均匀概率分布：概率与区间长度成比例（每个相同长度的子区间，他们的概率是相等的）

均匀概率密度函数公式：　注意：f(x)不是 x 的概率，这只是密度函数

$f(x) = \begin{cases} 1/(b-a) & a<=x<=b \\ 0 & \text{其他} \end{cases}$　　就如定义，考虑的是特定区间内发生的概率

但这个不是概率，是密度函数，概率靠下面的面积来算

数学期望：$E(x) = (a+b) / 2$　　a 为随机变量所能取的最小值

方差：　$Var(x) = (b-a)^2 / 12$　　b 为随机变量所能取的最大值

在给定区间$[x_1, x_2]$上取值的概率，也就是在区间$[x_1, x_2]$上概率密度函数 f(x)曲线下的面积

2.正态概率分布：最重要的概率分布，广泛的应用，如身高体重，成绩，降雨量

正态概率密度函数：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ-均值（期望）　π=3.14159　e=2.71828

σ-标准差

特点：1.最高点在均值 μ 处达到，并且均值 μ 还是中位数和众数

2.均值 μ 与标准差 σ 确定了正态分布的位置与形状

3.均值可以是任何值，是 0 那就是以 y 轴对称，是-1 那就是以 x=-1 的轴对称，必定左右对称，偏度为 0

4.标准差决定了曲线的宽度和平坦程度，越大越宽，越平坦，表示数据具有更大的变异性

5.有 68.3%的值在均值加减 1 个标准差的范围

有 95.4%的值在均值加减 2 个标准差的范围

有 99.7%的值在均值加减 3 个标准差的范围

注意：这是密度函数，说白了没用。。。。。。。，只是画图

标准正态概率分布 ----- 特殊的正态概率分布，均值 μ=0，标准差为 1

标准正态概率密度函数：

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{\left(-\frac{x^2}{2}\right)}$$

有一张数学用表，里面是累计概率，可以得出指定 z 的累计面积概率，或者给定面积或概率，确定相关的 z

要找 z<=1.25 的概率，那么找到 z=1.2 的行，然后是 z=0.05 的列，这个参数就是累计概率

正态概率分布的概率：（这是概率的算法！！！！）

我们要转换成标准正态概率分布： z = (x-μ)/σ　μ-均值 σ-标准差 x-随机变量

例如，假设均值 μ=10，标准差 σ=2，我们想计算 x 在 10-14 上取值的概率

那么当 x = 10 时，z = 0，当 x = 14 时，z = 2，

等价于标准正态概率分布随机变量 z 在 1-2 取值的概率，

P(0<=z<=2) = P(z<=2) - P(z<=0)

3.使用正态分布来模拟二项概率分布：

前提是二项概率分布的 n 很大，并且 np>=5 且 n(1-p)>=5

于是正态分布的：　μ = np

σ = √(n*p*(1-p)

二项概率分布是取特定随机变量的数值，如 x=10，但是正态分布是区间的概率

故正态分布可以取 9.5<=x<=10.5，其中 0.5 就是连续型校正因子

4.指数概率分布：指数分布常被用于描述服务时间

例如：到达某洗车处的两辆车的时间间隔

装载一辆卡车的所需时间

高速公路上两起重大事故发生地之间的距离

密度函数：

$$f(x)=\frac{1}{\mu}e^{-x/\mu}$$　x>=0　μ-均值，期望值

累计概率：

$$P(x<=x_0)=1-e^{-x_0/\mu}$$

注意：

1.标准差与均值一致

2.与泊松分布联系紧密：就是说适用于泊松分布，也就可以适用指数分布

假设每小时到达洗车店的车辆均值是 10

泊松分布 --- 每小时有 x 辆车到达洗车店的概率

指数分布 --- 两车到达的时间间隔概率

☆☆☆☆☆☆☆☆☆☆☆　基本术语　☆☆☆☆☆☆☆☆☆☆☆

统计资料：　指数据事实，比如具体的数值，平均数，中间数，最大值等

数据：　为了描述和解释所搜集、分析、汇总的事实和数字

数据集：　将用于特定研究而搜集的所有数据称为研究的数据集

例如参加 WTO 的 60 个国家信息的数据集

个体：　　指搜集数据的实体
　　例如每个国家
变量：　　个体中所感兴趣的那些特征
　　例如 WTO 身份，人均 DGP，惠誉评级等
观测值：　　对每个变量收集的测量值，得到数据的该个体测量值集合
　　60 个个体就该有 60 个测量值

数据项总数 = 个体的个数 * 变量的个数

测量尺度：取得并记录一个特定变量的数据
　　名义尺度 ----- 可以使用数值代码，比如 1 代表成员国，2 代表观测国
　　顺序尺度 ----- 数据的顺序，等级意义明确，如 AAA,B+,C，或者是考试分数的排名
　　间隔尺度 ----- 永远是数值型的，可以度量数值间的间隔，如屏幕的尺寸
　　比率尺度 ----- 含间隔尺度，并且两个数值之比是有意义的，如房价的环比，学生的分数

总体：　　在一个特定研究中所有感兴趣的个体组成的集合
样本：　　样本是总体的一个子集
普查：　　对总体全部数据的调查过程称为普查
抽样调查：　　搜集样本数据的调查过程
统计推断：　　利用样本数据对总体特征进行估计和假设检验

样本统计量：　计算的度量数据来自样本
总体参数：　　计算的度量数据来自总体
点估计量：　　样本统计量被称为相应总体参数的点估计量

大数据：　　更大和更为复杂的数据集，容量，速度，种类 3 个特点
数据挖掘：　　研究从非常大的数据库中开发有用的决策信息的方法，利用了统计学，数学，计算机科学
　　优化点是自动性，预测性

概率：　　对事件发生的可能性的数值度量
随机试验：　　是一个过程，它所产生的结果是完全确定的，但每次试验的结果是偶然的，例如抛硬币
样本空间：　　所有试验结果所组成的一个集合
事件：　　样本点的集合
样本点：　　一种特定的试验结果

描述统计：　　将数据以表格，图形或者数值形式汇总的统计方法
数据可视化：　用于汇总和表述一个数据集信息的图形显示，重点在图形，将数据可视化