深入浅出统计学笔记

第一章：信息图形化

统计 statistics：通过某种有意义的方式对原始事实和数据进行提炼，使得仅仅通过观察原始数据无法立即水落石出的一些理念得以昭示。

频数 frequency：表示在一个特定区间内的统计对象的数目。

饼图 pie chart：能够很好的体现基本比例。

条形图 bar chart：更灵活，相比较饼图更精确。

直方图 histogram：适合体现分组数值型数据。

直方图与条形图区别：直方图每个长方形没有间隔，直方图每个长方形面积与 frequency 成正比。（在统计学中，frequency-频数；relative frequency-频率。频数体现的是次数，频率体现的是比例；在物理学中，译为频率，意思是单位时间内完成周期性变化的次数，体现的也是次数。）
https://www.zhihu.com/question/26894953/answer/97516583 直方图一般用来描述等距数据或等比数据；柱形图一般用来描述称名数据或顺序数据。直观上，直方图矩形之间是衔接在一起的，表示数据间的数学关系；柱形图则留有空隙，表示仅作为两个或多个不同的类，而不具有数学相关性质
https://www.zhihu.com/question/26894953/answer/95721940 可以这么说，直方图的 Y 轴是频率，柱形图的 Y 轴可以是数值。直方图其实是柱形图的一种。用 Excel 做直方图就要先分组，计算频率，然后以频率为变量做柱形图

要 点

- **频数密度**指的是分组数据中的频数的密集度。计算方法如下：

  $$频数密度 = \frac{频数}{组距}$$

- **直方图**是一种专门用于体现分组数据的图形。它看起来很像条形图，但每条长方形的高度等于频数密度——而不是频数。
- 绘制直方图时，每个长方形的宽度与其分组宽度（“组距”）成正比例。长方形按照连续的数字标度绘制。
- 直方图中的每个组的频数通过长方形面积求出。
- 直方图的长方形之间没有间隔。

数值型数据（定量数据） numeric data：涉及数字和数量； 类别型数据（定性数据） categorical data：涉及表述，描述性质和特征。

第二章：集中趋势的量度（均值，中位数，众数）

向左/右偏斜 skewed tothe left/right

均值 mean： $\mu = \sum x / n$

中位数 median：奇数个数值，中位数的位置（n+1）/ 2，偶数个数值，两个中位数求和取均值，这两个中间数位于（n+1）/ 2 两侧

众数 mode：必须是数据集中的一个数值，而且是最频繁出现的数值。众数是唯一能用于类别数据的平均数类型

第三章：分散性与变异性的量度

全距（极差） range：用于量度数据集分散程度的一种方法，最大值/上界 upper board - 最小值/下界 lower board。全距仅仅描述数据的宽度，并未描述数据在上下界之间的分布形态

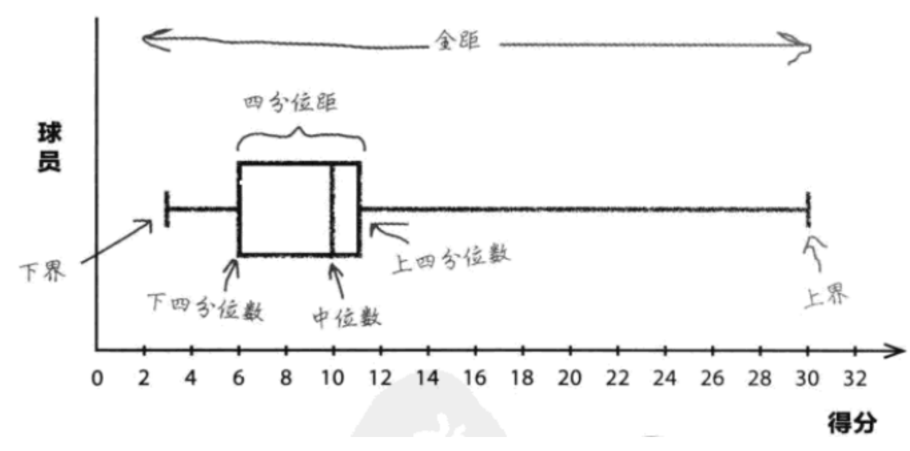
四分位数 quartile：将数据四等分的几个数值。最大的四分位数称为上四分位数 upper quartile，最小的四分位数称为下四分位数 lower quartile。中间的四分位数即为中位数 median

四分位距 interquartile range(IQR)：上四分位数 - 下四分位数

下四分位数 lower quartile：n/4，如果为整数，取这个位置和下一个位置的均值；如果为小数，向上取整表示下四分位数的位置

上四分位数 upper quartile：3n/4，如果为整数，取这个位置和下一个位置的均值；如果为小数，向上取整表示下四分位数的位置

箱线图 box and whisker（或者箱型图 box plot）：箱线图显示数据的全距、四分位距以及中位数



方差 variance：$s^2 = \sum(x-\mu)^2$ / n

$$s^2 = \frac{(x_1-M)^2+(x_2-M)^2+(x_3-M)^2+\cdots+(x_n-M)^2}{n}$$

$$S^2 = \frac{\sum x^2}{n} - \mu^2 \qquad S^2 = \frac{\sum x^2}{n} - \mu^2 \qquad S^2 = \frac{\sum x^2}{n} - \mu^2$$

(方差速算法：$s^2 = \sum x^2$ / n - μ^2)

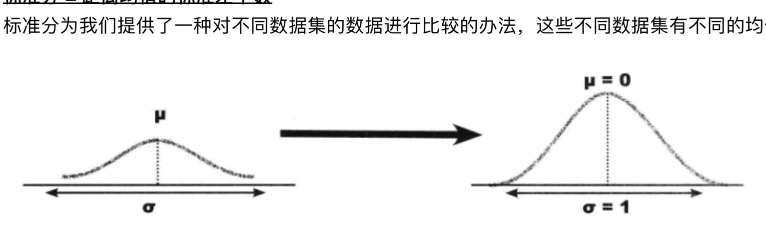
$$\begin{aligned}
\sigma^2 &= \frac{\sum_{i=1}^{N}(Xi-\mu)^2}{N} \\
&= \frac{\sum_{i=1}^{N}(Xi^2-2\mu * Xi+\mu^2)}{N} \\
&= \frac{\sum_{i=1}^{N}Xi^2}{N} - \frac{2\mu\sum_{i=1}^{N}Xi}{N} + \frac{\sum_{i=1}^{N}\mu^2}{N} \\
&= \frac{\sum_{i=1}^{N}Xi^2}{N} - 2\mu\frac{\sum_{i=1}^{N}Xi}{N} + \frac{\sum_{i=1}^{N}\mu^2}{N} \\
&= \frac{\sum_{i=1}^{N}Xi^2}{N} - 2\mu^2 + \mu^2 \\
&= \frac{\sum_{i=1}^{N}Xi^2}{N} - \mu^2 \\
&= \frac{\sum_{i=1}^{N}Xi^2}{N} - \frac{(\sum_{i=1}^{N}Xi)^2}{N^2}
\end{aligned}$$

（前述两种方式的转化过程）

标准差 standard deviation：σ = $\sqrt{variance}$

$$\sigma = \sqrt{S^2}$$

标准分 standard score(或者 z 分 z-score)：z =（x-μ）/ σ 标准分的作用是将几个数据集转换成一个理论上的新分布，这个分布的均值为 0，标准差为 1
标准分可以取任意值．这些值表示相对于均值的位置。正的 z 分表示数值高于均值，负的 z 分表示数值低于均值。若 z 分为 0，则数值等于均值本身。数值大小体现了数值与均值的距离
标准分 = 距离均值的标准差个数
标准分为我们提供了一种对不同数据集的数据进行比较的办法，这些不同数据集有不同的均值和标准差，通过标准分可以把这些数值视为来自同一个数据集或数据分布。

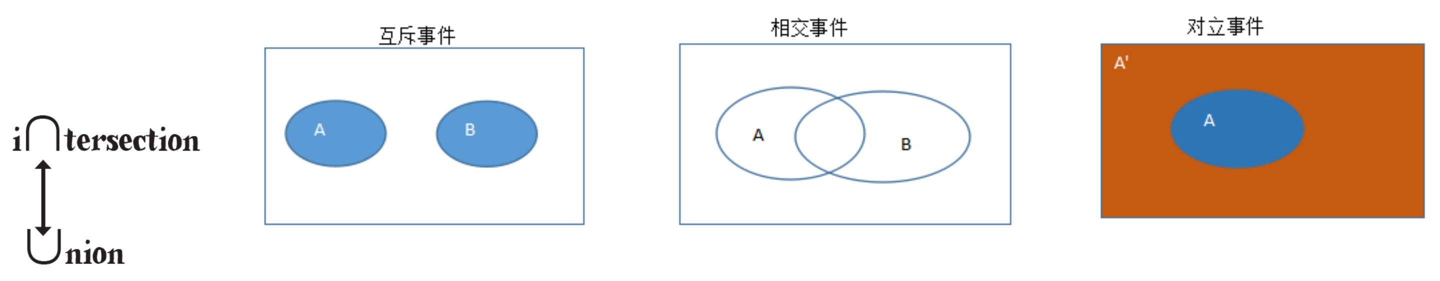


第四章：概率计算

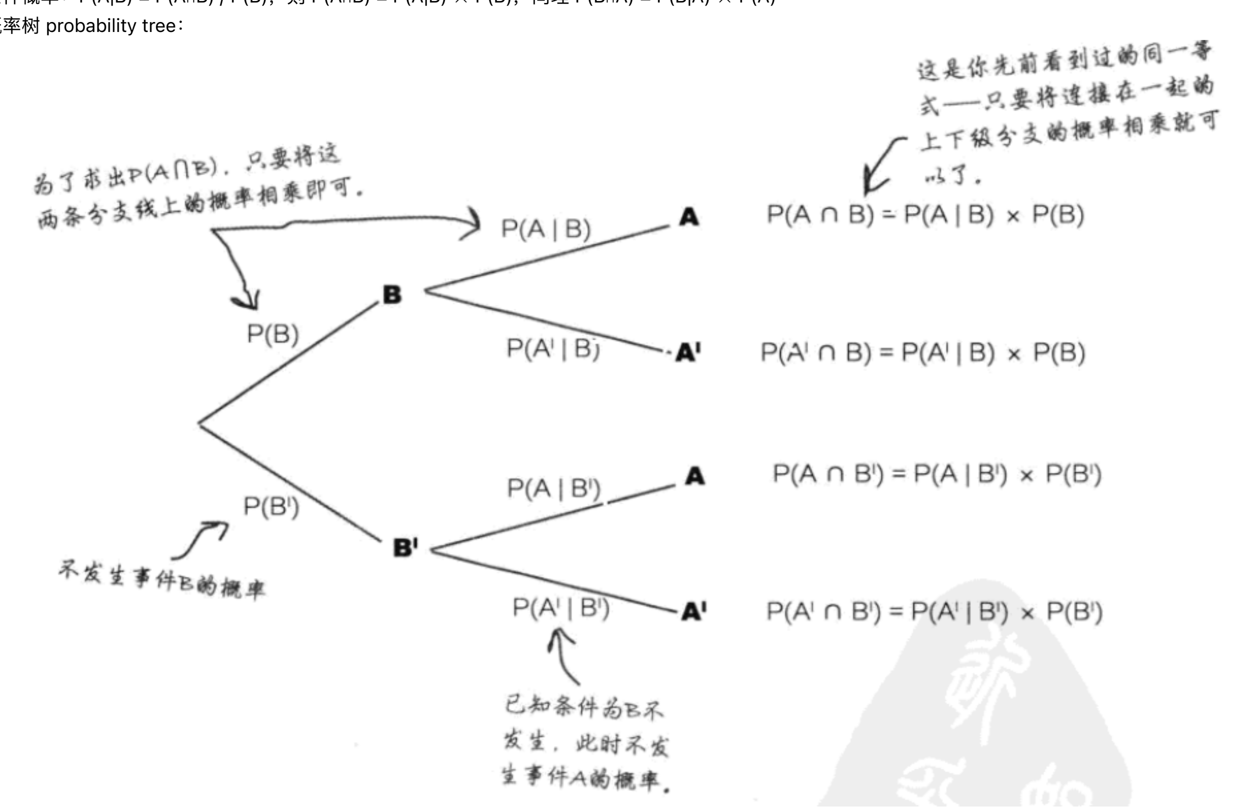
S 被称为概率空间 possibility space，或称样本空间 sample space，是表示所有可能结果的一种简便表示法。可能发生的事件都是 S 的子集

维恩图 Venn diagram

对立事件：“A 不发生”事件可以用 A' 表示。A' 被称为 A 的对立事件。A' 包含事件 A 所不包含的任何事件。 P(A') = 1 - P(A)
互斥事件：如果两个事件是互斥事件，则只有其中一个事件会发生，这两个事件不会同时发生
相交事件：如果两个事件相交，则这两个事件有可能同时发生。P(A∪B) = P(A) + P(B) - P(A∩B)



条件概率：P(A|B) = P(A∩B) / P(B)，则 P(A∩B) = P(A|B) × P(B)，同理 P(B∩A) = P(B|A) × P(A)
概率树 probability tree：



全概率公式：P(B) = P(A∩B) + P(A'∩B) = P(A) × P(B|A) + P(A') × P(B|A')

贝叶斯定理：已知P(A),P(B|A),P(B|A');求P(A|B). P(A|B) = P(A∩B) / P(B) = P(A)* P(B|A) / P(A)* P(B|A)+P(A')* P(B|A')

相关事件： 如果 P(A|B)不等于P(A)，就说事件A与事件B的概率相互影响。

独立事件：几个事件互相不影响。P(A|B)=P(A). 如果两个事件相互独立，则 P(A∩B)= P(A|B)P(B)=P(A)P(B)

第五章：离散概率分布的应用

离散型随机变量的期望：E(x)=∑xP(X=x)

离散型随机变量方差：Var(X)=E(x-μ)²=∑[(x-μ)²P(X=x)]

线性变换的通用公式： E(aX+b)=aE(X)+b; Var(aX+b)=a²Var(X)

E(aX+bY)=aE(X)+bE(Y); Var(aX+bY)=a²Var(X)+b²Var(Y)

E(aX-bY)=aE(X)-bE(Y); Var(aX-bY)=a²Var(X)+b²Var(Y)

第六章：排列与组合

排位方式与计算公式

求n个对象的可能排位方式的数目，n！=n*(n-1)(n-2)……3* 2* 1

如果是圆形排列，则可能的情况一共有 (n-1)！=(n-1)(n-2)……3* 2* 1

按照类型来排位：如果要为n个对象排位，其中包括一类对象有k个，另一类对象有j个，另一类对象有m个，则可能的排位情况有 n！/(k！ * j！ * m！)

排列

排列（考虑排序）：从一个较大（n个）对象群体中取出一定数目（r个）对象进行排序，并得出排序方式总数目：

组合（不考虑排序）：从一个群体选取几个对象，不考虑这几个对象的顺序，求出一共有多少种情况。

第七章：几何分布、二项分布、泊松分布

几何分布 X~Geo(p)

几何分布包含以下条件：

1、进行一系列相互独立的实验。 2、每一次实验成功概率为p，失败概率为1-p。 3、主要关注：为了取得第一次成功需要进行多少次实验。

几何分布的概率计算：
1、在第r次实验才成功的概率

2、至少需要r次实验才能成功的概率

3、需要实验r次或者不到r次就成功的概率

几何分布的期望和方差：

二项分布 X~B(n,p)

二项分布的条件：

```
1、进行一系列独立实验。
2、每一次实验成功概率为p，失败概率为1-p。
3、实验次数有限。
```

二项分布和几何分布情况一样，需要进行一系列实验，差别在于二项分布的关注点是获得成功的次数

二项分布概率计算：

二项分布期望和方差：

泊松分布X~Po(λ)

泊松分布条件：

```
1、单独事件在给定区间内随机，独立地发生，给定区间可以指时间或空间。
2、已知该区间内的事件平均发生次数（发生率），且为有限数值。该事件的平均发生次数用λ表示。
```

泊松分布概率计算： 求给定区间内，发生n次事件的概率：

X,Y都是独立随机变量，如果X~ Po(λ1)， Y~ Po(λ2)，则可以等效于X+Y~Po(λ1+λ2)。 如果X，Y都符合泊松分布，则X+Y也符合泊松分布。

特定条件下，泊松分布可以近似代替二项分布。 泊松分布的期望λ，方差λ。 二项分布的期望np，方差npq。当n特别大，q特别小。λ≈np。 所以二项分布可近似于X~Po(np)

第八章：正态分布（高斯分布）N(μ,σ²)

对于离散概率分布来说，我们关心的是取得一个特定数值的概率。对于连续概率分布来说，我们关心的是取得一个特定范围的概率

正态分布是连续数据的“理想”模型 如果一个连续随机变量X符合均值为μ，标准差为σ的正态分布，记做 N(μ,σ²) 正态分布计算三步法：

```
1、确定分布与范围。（确定 N(μ,σ²)中的均值，和标准差）
2、使其标准化。（求标准分 z=（x-μ）/σ）
```

3、查找概率。(用概率表查找概率) 概率表查到的是P（X<z）的概率。

第九章：正态分布的应用--超越正态

如果两组正态分布X~ N(μ1,σ1²),Y~ N(μ2,σ2²),X，Y为独立变量，则：

```
X+Y~N(μ1+μ2,σ1²+σ2²)
X-Y~N(μ1-μ2,σ1²+σ2²)
```

如果X~N(μ,σ²);则：

```
aX+b~(aμ+b,a²σ²)
```

如果X1，X2，X3为一系列独立的连续变量，且都满足正态分布X~N(μ,σ²)则：

```
X1+X2+X3……Xn~N(nμ,nσ²)
```

用正态分布近似代替二项分布

在某些特定情况下，可以用正态分布近似代替二项分布，如果二项分布X~B(n,p),且np>5,nq>5;则可以用正态分布X~N(np,npq)，近似代替X。 **但是还需要对正态分布进行连续性修正，才能保证得到正确的结果**

关于连续性修正：

```
≤型：如果用正态分布求P(X≤a)，实际是求P(X<a+0.5)
≥型：如果用正态分布求P(X≥a)，实际是求P(X>a-0.5)
介于型：如果用正态分布求P(a≤X≤b)，实际是求P(a-0.5<X<b+0.5)
```

用正态分布近似代替泊松分布

如果泊松分布X~ Po(λ)，λ>15,则可以用正态分布X~ N(λ,λ)代替。**需要进行连续性修正**