# 目录

# Sets

Set 的定义就是不重复的物体的合集. 这个物体可以是任何真实的/抽象的物品，如某个人，某个数，或者某个颜色.

并且，我们可以用多个 Sets 进行计算，来得到一个新的 Set

还有一些 Set 之间关系的表示符号：

- ∅：空集，当一个集合没有任何元素
- ∈：属于，也就是某个元素是存在于某个集合中：$1 \in \{1,3,5\}$
- ⊆：某个集合是另一个集合的子集：$\{1,9\} \subseteq \{1,3,9,11\}, \{2,4\} \nsubseteq \{1,3,9,11\}$
- Ω: Sample Space：一个特殊的字符，表示所有可能的组合：Coin Flip：$\Omega = \{\mathrm{H}, \mathrm{T}\}$
- Event：某一件事的结果，是 Sample Space 的子集合. 如扔骰子得到双数：$\{2,4,6\} \subseteq \{1,2,3,4,5,6\} \subseteq \Omega$
- ∪：两个集合的合集：$\{1,2\} \cup \{2,3,4\} = \{1,2,3,4\}$ 如果集合 A 和集合 B 是两个 Event，那么就表示是元素既有可能是 Event A 或者 Event B，或者两个都是：

$$A = \{1,3,5\}\ toss\ a\ odd\ number$$
$$B = \{1,2,3\}\ toss\ a\ number\ less\ than\ 3$$
$$A \cup B = \{1,2,3,5\}$$

- ∩：两个集合的并集:$\{1,2,3,6\} \cap \{2,4,6,8\} = \{2,6\}$对于 Event 也是同理，如果两个 Event 的并集，那么最终结果是需要同时满足两个 Event，如上面的$A \cap B = \{1,3\}$ 如果两个并

集是空集，那么我们就可以说这两个集合是 disjoint.

- $A^c$：表示的是集合 A 的补集，也就是在所有可能中，除 A 集合之外的结果，我们可以用它做集合相减的操作：$A - B = A \cap B^C$，并且这个也是遵守 DeMorgan's Law: $(A \cup B)^C = A^C \cap B^C$，因为我们可以把这个看作是~$A\ or\ !A$
- $|A|$：表示的是一个集合中有多少元素
- ×：表示做实验 2 次，比如我们做实验两次$\Omega \times \Omega = \{(x, y): x \in \Omega, \mathrm{y} \in \Omega\}$
- $\Omega^n$：表示做实验多次，如做实验 10 次，$\Omega^{10}$

# Probability

概率表示的是某个事件在所有可能的事件($\Omega$)中发生的概率，且:

$$P(\Omega) = 1$$
$$P(A \cup B) = P(A) + P(B)\ when\ A \cap B = \emptyset$$

如$P(A)$表示的就是 A 事件发生的概率

当$\Omega$中的可能情况是有限的，且每个事件的概率都相等，我们可以这么计算$P(A) = \frac{|A|}{|\Omega|}$

## Permutations

Permutations 指一个有顺序的 set,比如$\{1,2,3\}$和$\{3,2,1\}$是两个不同的 permutations. 对于一个 n 大小的 set，一共有$n!$种排列

## Conditional Probability

这个的意思就是当我们已经知道了事件 B 已经发生，另一件事 A 发生的概率，写成$P(A|B)$

具体的计算方法是：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

当然，这个也可以反过来求$P(A \cap B) = P(A|B)P(B)$，虽然看起来$P(A|B)$很少见，但是其实比如袋子 B 里有 3 个白石头就可以写作已知选中了袋子 B，然后选中白石头的概率就是

例 1：扔两个骰子，事件 A 是两个骰子的总和是 5，事件 B 是至少其中一个骰子的结果是 2. 求$P(A|B)$

首先，我们知道有 36 种情况, 11 种情况是满足 B 事件但是其中满足 A 事件的也就只有两件$\{2,3\}$和$\{3,2\}$，所以$P(A \cap B) = \frac{2}{36}$，$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{36} * \frac{36}{11} = \frac{2}{11}$

例子 2：我们还能用这个化简一些情况，比如一个罐子里有红蓝石头各 20 个，求拿出两个红石头的概率是多少．当然，我们可以直接求概率：$P(R \cap R) = \frac{1}{2} * \frac{19}{39}$

如果用上面的公式，就可以变成：$P(R \cap R) = P(R|R) * P(R)$，$P(R) = \frac{1}{2}$，而$P(R|R)$的意思是已知抽出红石头，再抽出红石头的概率，所以就是$\frac{19}{39}$，两个一乘就是上面的概率

例子 3：有两个人，已知其中一个是男的，求两个都是男的概率是多少?

按照直觉，因为我们已经知道了其中一个人是男的了，剩下的一个有$\frac{1}{2}$可能是男的，所以两个都是男生的概率应该是$\frac{1}{2}$.

用上面的公式来测试一下，设两个都是男生的条件是 Q, S:一个是男生

$$P(Q|S) = \frac{P(Q \cap S)}{P(S)} = \frac{1/4}{3/4} = \frac{1}{3}$$

$P(S)$是四种组合中只要出现一个男生就满足，所以是$\frac{3}{4}$

$P(Q \cap S)$是既要满足两个男孩还要满足至少有一个男孩，所以概率只有$\frac{1}{4}$

最终得到$\frac{1}{3}$，虽然有点反直觉，但却是是这样的(但是前提是你不知道到底哪个人是男生)

有时候，我们要求一个概率，但我们只知道在某些条件下的概率的话，我们就可以把所有可能的条件概率乘以这个条件所需的概率加一起，得到最终的概率.

例：疯牛病的发病概率是 0.02，真阳性的概率是 0.7(牛得病,且测出来得病)，假阳性的概率是 0.1(牛没得病但测出得病)，求测试出阳性的概率是多少：

设得病事件是$P(b)$，阳性是$P(p)$

$$P(p) = P(p|b)P(b) + P(p|b^c)P(b^c) = 0.7 * 0.02 + 0.1 * 0.98$$

上面之所以要求两个是因为对于$P(p)$来说，有两种情况：第一种是真阳性，第二种是假阳性从此能看出，对于测试来说，这个假阳性反而会更有可能影响最终测试结果，因为没得病的概率更高，导致虽然假阳性的概率低，但是次数多，所以更有可能影响结果.

## Bayes Rule

有时候，我们只知道结果，但不知道过程，就可以用这个公式

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

例 1：有两个罐子，一个有 4 黑 3 白石头，另一个有 2 黑 2 白，如果别人帮我选了一个罐子和一个石头，是黑色的，问选中罐子 1 的概率是多少?

就可以用上面的公式：$P(u1|B) = \frac{P(B|u1)P(u1)}{P(B)} = \frac{\frac{4}{7}*\frac{1}{2}}{\frac{1}{2}*\frac{4}{7}+\frac{1}{2}*\frac{1}{2}}$

例 2:

有三个机器, A 机器每小时产 10 个零件，10%是坏的, B 机器每小时产 20 个零件，5%是坏的, C 机器每小时产 30 个零件，1%是坏的. 所有的零件都汇聚在一个箱子里，选了一个零

件，发现这个是坏的，这个零件来自 B 机器的概率是多少?

首先，我们需要先知道随便选一个零件，这个零件来自各个机器的概率是多少，每个机器产坏零件的概率是多少:

$$P(A) = \frac{1}{6}, \qquad P(F|A) = 0.1$$
$$P(B) = \frac{2}{6}, \qquad P(F|B) = 0.05$$
$$P(C) = \frac{3}{6}, \qquad P(F|C) = 0.01$$

根据 Bayes Rule:

$$P(B|F) = \frac{P(F|B)P(B)}{P(F)} = \frac{0.05 * \frac{2}{6}}{0.1 * \frac{1}{6} + 0.05 * \frac{2}{6} + 0.01 * \frac{3}{6}} \approx 0.435$$

# Independence

如果两个事件是不相关的，那么一个事件的出现就不会影响到下个事件出现的概率，写做

$$P(A|B) = P(A)$$

因为上面的可以转化成$P(A \cap B)/P(B)$，所以

$$P(A \cap B) = P(A)P(B)$$

又因为$P(A \cap B)$里面的 A 和 B 可以交换，所以又可以写成

$$\frac{P(B \cap A)}{P(A)} = P(A|B) = P(B|A)$$

上面这些之所以相等是因为 AB 事件完全不相关. (以上不确定)

并且，当$A, B$不相关，$P(A^c|B) = P(A^c)$，因为$P(A^c|B) = 1 - P(A|B) = 1 - P(A) = P(A^c)$

如果我们想证明两个事件是否相关的话，可以利用上面这些特性，如果等式两边的数不相等，就说明这两个事件是相关的.

例：有个电源和辅助电源，主电源有10%的几率失效，主电源工作时副电源也有10%的概率失效，但如果主电源失效了，副电源会有更高概率失效15%．求主副电源是否相关．

设主电源失效概率为$P(M)$，副电源失效是$P(A)$，$P(A|M^c) = 0.1$，$P(A|M) = 0.15$，

$0.1 \neq 0.15$，所以两个事件相关．

## Tree Diagram

用树状图把所有的可能性给画出来，这样能更好的看见某一件事发生的概率：比如有两个箱子，一个箱子里有数字$\{1,3,5\}$，另一个箱子里有数字$\{2,4\}$，我们先选一个箱子再选一个数字．所以我们选择到数字 2 的概率是多少?

首先，我们需要先随机选择一个箱子，但选择到正确箱子的概率只有$\frac{1}{2}$，所以我们需要先乘上一个$\frac{1}{2}$，接着再选数字：由于有两个数字但只有一个正确的，所以选到正确数字的概率也是$\frac{1}{2}$．但只有两个事件同时都满足才能选到数字 2，或者说后面的这个事件是基于前面的这个事件的，所以要把两个概率乘起来$P(S \cap R) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$

## Sampling without Replacement

在上面这个事件，上个事件发生过后并不会影响下一件事发生的概率，但是在这种情况下不一样：如果我们从箱子里取出一个球，但不放回去，那么下次再次抽取到这个球的概率就会变低．在这种情况下, Tree diagram 也需要根据抽取的结果进行更改，比如从一个有红蓝各 20 个球的箱子里抽两个红球，他的概率是$P(R \cap R) = \frac{20}{40} * \frac{19}{39}$，但如果拿出来后再放回去的概率就是$P(R \cap R) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$

# Random Variable

随机变量是一种用来表示随机现象各种结果的数值函数. 在统计学中，随机变量可以用来量化和分析不确定的数据和事件.

如:

扔两个骰子，设$x$为两个骰子的结果的和，问$\Pr(\{x = 4\})$

用来画表，我们可以得到:

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |

从上面的表可以看到，得到的结果有 36 种可能，其中有 3 个是等于 4，所以概率是$\frac{3}{36}$

在这里，$x$就是$Random\ Variable$

如果我们把所有的 Random Variable 的概率写下来，会是$\frac{1}{36}, \frac{2}{36}, \frac{3}{36} \ldots, \frac{1}{36}$，如果我们画出图的话，就可以看到是 11 根离散的柱子，离散是因为两个骰子的和只会是整数

## Probability Distribution Function

$$F(a) = \Pr(\{x \leq a\})$$

还是取之前两个骰子的和，如果我们要算$F(4)$，就是所有小于等于 4 的组合，也就是 6 种，

概率也是$\frac{6}{36}$，也就是说，这个$a$越大，概率越大，从$F(1)=0$ 到 $F(12)=1$

## Common Distribution

除了我们得出的概率分布，还有一些预定义的概率分布，其中最常见的就是

**Bernoulli Distribution:**

这个的特点是我们会定义一个数值$p$，$f(1)=p, f(0)=1-p$更直白的说其实就是只有两种结果, 0 和 1，得到 1 的概率是那么多，那么得到 0 的概率就是剩下的

$$f(x)=p^x(1-p)^{1-x}\begin{cases} p\ if\ x=1 \\ 1-p\ if\ x=0 \end{cases}$$

**Binomial Distribution:**

这个是我们不停重复 Bernoulli trail 最后得到的概率，并且每次实验互不影响，且顺序无关

比如我们翻硬币$n$次，求$k$次正面的概率

首先要求有多少种抽到的可能:

$$\binom{n}{k}=\frac{n!}{(n-k)!\,k!}$$

这个表示的是如果我们要从 n 个物品中抽出 k 个物品，一共有多少种组合(顺序不重要)

而 Binomial Distribution 的计算方法是

$$\binom{n}{k}p^k(1-p)^{n-k}$$

如果我们要扔一个正面概率是$0.6$的硬币 10 次，问出 4 次正面的概率是多少?

$$n=10$$
$$k=4$$
$$p=0.6$$
$$\Pr(\#head=4)=\binom{n}{k}0.6^k*0.4^{n-k}$$

我们用这种计算方式而不是直接用$0.6^4$是因为后者是连出 4 次的概率，而前者，也就是这个公式是指 10 次里面出 4 次，且不关心顺序的概率.

**Geometric distribution:**

几何分布是指在一系列独立的伯努利试验中，要达到第一次成功所需要的试验次数的概率.

伯努利试验是指每次试验只有两种可能的结果，比如正面或反面，成功或失败，等等．每次试验的成功概率是固定的，不受前后试验的影响.

使用公式求第$k$次是第一次成功的概率是:

$$\Pr(k) = (1-p)^{k-1}p$$

其实就是前面连着出现$k-1$次 “失败”的事件，然后再接着出现一次 “成功”的事件

## Continuous Random Variable

在之前，我们都是在计算离散的概率，要是我们想要知道一个不是离散的概率呢？例如我们想要转一个圆盘，求停在某处区域的概率是多少?

$$\Pr(a \le x \le b) = \int_a^b f(x)\,dx$$

改一下 a 和 b，就得到下面这个

$$\Pr(a-\varepsilon \le x \le a+\varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,dx$$

有些时候，求积分可能不好求，我们可以用近似法：用两个长方形近似概率的面积.

$$2 * \varepsilon * f(a)$$

因为宽度是两个$\varepsilon$，乘上长方形的高$f(a)$就是近似的概率

如果我们把它写成 CDF 的形式(概率累计)，那么$F(a) = \Pr(x \le a)$这个概率公式

$$F\left(\lim_{a\to-\infty} a\right) = 0$$

$$F\left(\lim_{a\to\infty} a\right) = 1$$

而且，在实际计算的时候，我们根本不需要算积分，直接用 CDF 就行(如果我们已经知道这两个 cdf 概率)

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

**PDF**

表示的是 probability density Function，也就是上面的，我们通过积分一段区域来获得这片区域的概率，但需要知道的是，对于 PDF 来说，不可以有概率是副的.

$$\begin{cases} \frac{1}{2} - x^2 & for\ x \in [-1,1] \\ 0 & otherwhere \end{cases}$$

上面这种就不是一个 valid pdf，因为如果$x = 1$的时候，就会是负数

$$\begin{cases} \sin(\mathrm{x}) \text{ if } \mathrm{x} \in [\frac{\pi}{2}, \pi] \\ 0 & otherwhere \end{cases}$$

上面这种事一个 valid pdf，因为没有区域是负数，积分出来也正好是 1

**Uniform Distribution**

这个表示的是在$X$区间$[a,b]$中概率分布是相等的. 也就是$\frac{1}{b-a}$，比如从 ab 分别是 01，那么概率都是 0.1，累计概率$F(x)$则是一个线性上升的函数

如果我们要算累计概率就是

$$F(a) = \int_\alpha^a \frac{1}{\beta - \alpha}\,dx = \frac{a - \alpha}{\beta - \alpha}$$

例：公交车来的概率在$[3,7]$分钟区间内均匀分布，如果我只想等 5 分钟，那么等到的概率是多少?

$$F(5) = \frac{5-7}{3-7} = \frac{1}{2}$$

**Exponential distribution**

这种概率的 pdf 是$\lambda e^{-\lambda x}$, cdf 是$1 - e^{-\lambda x}$，写作$x \sim Exp(\lambda)$

如果画上图，就可以知道如果$\lambda$越大，事件发生的越快(pdf 下降的越快)

意思就是这种事件在未来是随机发生的(第一次)，就像连续版的 Geometric distribution

比如我们在做爆米花，第一个爆米花在第 5 秒之后爆开的概率是多少?

$$x \sim Exp\left(\lambda = \frac{1}{2} second\right)$$

$$1 - F(5) = 1 - \left(1 - e^{-\frac{1}{2}*5}\right) = e^{-\frac{5}{2}} = 0.082$$

## Expectations

中文来讲，是期望，也就是如果有一个 Random Variable，在无数次实验过后，出现次数最多的结果是多少.

或者还有平均期望(Mean Expectation)，也就是实验无数次后结果的平均数. 换句话说，我们也可以把这个平均期望看做是这个概率图的中心.

如抽三个数，抽到$-1$是$0.2$，抽到$1$是$0.2$，抽到$0$是$0.6$的概率，则重心,/平均期望，就是 0

$$E[X] = \sum_i a_i P(X = a_i)$$

$$E[X] = \int_{-\infty}^{\infty} a_x P(X = a_x)\, dx$$

对于 Uniform Distribution，期望比较好算，由于$P(X = x) = \frac{1}{b-a}$，带入上面积分公式最后得到$E[X] = \frac{a+b}{2}$

对于伯努利分布，$P(X=x)=p^x(1-p)^{1-x}, x\ in\ \{0,1\}$，由于 x 在 0-1，并且是离散的，所以用 sigma 计算：$\sum_x xP(X=x)=0+1*P(X=1)=p$

$$E[X]=p$$

对于几何分布，$P(X=x)=p(1-p)^{x-1}$，$E[X]=\sum_{k=1}^{\infty}kp(1-p)^{k-1}=\frac{1}{p}$

Exponential distribution 的$P(X=x)=\lambda e^{-\lambda x}$，$E[X]=\frac{1}{\lambda}$

别看上面这些公式比较复杂，其实很好理解，就拿几何分布来说，我们是要找第一次成功的平均值，所以成功率$p$越高，这个值也就越小，如果$p=1$,也就是$100\%$成功，所以平均值也就是 1 次成功

如果我们修改了这个 Random Variable 的值，那计算结果的时候，概率不变，但是数值会变:

$$E[g(X)]=\sum_i g(a_i)P(X=a_i)$$

如:

$$X\sim Ber(p),\qquad compute\ \ E[2^X]$$

$$E[2^X]=\sum 2^XP(X=x)=2^0P(x=0)+2^1P(x=1)=1*(1-p)+2p=1-p+2p$$
$$=1+p$$

例 2 (期望的线性性质):

$$E[aX+bY]=aE[X]+bE[Y]$$
$$\sum_{x,y}(aX+bY)P(x,y)=\sum_{x,y}aXP(x,y)+bYP(x,y)$$

$$= \sum_{x,y} aXP(x,y) + \sum_{x,y} bYP(x,y)$$

简单来说，就是期望是可以拆开的，比如扔 6 个骰子得出的和的期望是等于 6 乘以一个骰子得出的值的期望．但需要注意的是，如果要乘的话，只能拆开期望乘以常数得到的期望．不能拆开两个期望相乘得到的新期望．

或者，非线性组合也不行：如两个骰子大减小得到的期望也不能拆开，因为大减小或者$abs()$让这个 Random Variable 不再线性．

$$E[X] = E[abs(X_1 - X_2)]$$

如果不是大减小，那么最终期望就是$3.5 - 3.5 = 0$，但如果加上 abs，那我们就得一个一个求概率和结果用∑来算了

例 3:

扔 10 个骰子然后相加，请问期望是多少?

$$E[10x] = 10E[x] = 10\sum_{x=1}^{6} x * \frac{1}{6} = 10 * \frac{1}{6} * \sum_{x=1}^{6} x$$

当我们要求非线性的期望的时候，就不一样了:

有一个正方形的房子，他的边长是10~20米(均匀分布)，求期望的面积是多少?

对于期望边长来说，$E[X] = \frac{a+b}{2} = 15$，而$E[X^2] \neq 15^2$，因为面积增长并不是线性的，我们需要把对应的边长乘方，再乘上对应的概率再积分:

$$\int_{10}^{20} x^2 * \frac{1}{b-a} dx \rightarrow \int_{10}^{20} x^2 * \frac{1}{10} dx$$

## Variance

方差表示的是表示的是一个随机变量的扩散大小

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$
$$Var \geq 0, \qquad E[X^2] \geq (E[X])^2$$

而标准方差是$std(X) = \sqrt{Var(X)} = \sigma$

如果我们想计算 Variance，我们可以

$$\mu = \int_{-\infty}^{\infty} a * f(a)\, da$$

其中$f(a)$是事件在$a$发生的概率.

例:

求 Uniform distribution 的 Variance:

$$\mu = \int_{a}^{b} x * \frac{x-a}{b-a} dx = \frac{b-a}{2}$$

$$\mu = 0(1-p) + 1 * p = p$$
$$Var[X] = (0-p)^2(1-p) + (1-p)^2 * p = p * (1-p)$$

绿色表示$f(a_i)$

蓝色表示$(a_i - \mu)^2$

| 分布 | 密度函数 | 数学期望 | 方差 |
| --- | --- | --- | --- |
| Bernoulli Distribution<br>0-1 分布 | $P(X = k) = p^k(1-p)^{1-k}$ | $p$ | $p(1-p)$ |
| Binomial Distribution<br>二项分布 | $P(X = k) = C_k^n p^k(1 - p)^{n-k}$ | $np$ | $np(1-p)$ |
| Geometric distribution:<br>几何分布 | $\mathrm{P}(X = k) = (1-p)^{k-1}p$ | $\frac{1}{p}$ | $\frac{1-p}{p^2}$ |

| | | | |
|---|---|---|---|
| Uniform Distribution<br>均匀分布 | $f(x)=\begin{cases}\frac{1}{b-a}, a<x<b\\ 0, other\ wise\end{cases}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| Exponential distribution<br>指数分布 | $f(x)=\begin{cases}\lambda e^{-\lambda x}, x>0\\ 0, otherwise\end{cases}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| Normal Distribution<br>正态分布 | $f(x\|\mu,\sigma^2)$<br>$=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ | $\mu$ | $\sigma^2$ |

# Joint Probability

Joint probability（联合概率）是指在概率论和统计学中，两个或多个随机事件同时发生的概率. 如果有两个事件 A 和 B，它们的联合概率表示为 P(A ∩ B)，即事件 A 和事件 B 同时发生的概率. 联合概率可以用于描述多个随机变量之间的关系，以及它们同时取特定值的可能性.

$$\sum_i \sum_j f_{xy}(a,b)=1$$

但需要注意的是，$P(A=a,B=b)\leq P(A=a)$ 因为前者需要的条件更多，不仅发生 a 还要发生 b，所以概率一定小于后者

看起来很复杂，但来举个例子吧：

某个视频网站调查了男女喜欢不同电影的数量

| | Male | Female | Total |
|---|---|---|---|
| Game of Thrones | 80 | 120 | 200 |
| West World | 100 | 25 | 125 |
| Other | 50 | 125 | 175 |

| | | | |
|---|---|---|---|
| Total | 230 | 270 | 500 |

当然上面的只还是一个统计，当我们把数据除以整体(500)就成了概率：

| | Male | Female | Total | $P(Show\|Female)$ |
|---|---|---|---|---|
| Game of Thrones | 0.16 | 0.24 | 0.4 | 0.444 |
| West World | 0.1 | 0.05 | 0.25 | 0.093 |
| Other | 0.1 | 0.25 | 0.35 | 0.464 |
| Total | 0.46 | 0.54 | 1 | 1 |

其中每个单元格都是一个联合概率：如喜欢权力的游戏的女性观众的概率是 0.24

右边最后一列和下面最后一行(总和)我们称作 Marginal Probability，它忽略了其中一个条件，只统计了一个条件下的概率(如男性观众的概率，或看权力的游戏的观众的概率)

如果我们想求两个联合概率，我们可以用这个公式:$P(A \cup B) = P(A) + P(B) - P(A \cap B)$，也就是两个编辑概率相加再去掉重合的概率

**条件概率**

如果我们想求条件概率，我们可以用贝叶斯定理：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

如：一个女性订阅了这个视频网站，它是权力的游戏这个电影的粉丝的概率是什么?

在这里，已知条件 B 就是他是女性，未知条件就是权力的游戏的概率，$P(B) = 0.54$ 也就是边际概率，$P(A \cap B) = 0.24$ 也就是女性喜欢权力的游戏的概率，两个相除就是最终结果

如果我们把所有的这种概率都算出来，也就是已知是女性，那这位女性喜欢某个电影的概率，$P(Show|Female)$，我们就得到了条件概率分布-Conditional Probability Distribution

用上面这个数据，我们可以得到两个东西是否相关:$If\ independent: P(A|B) = P(A)$

$$P(West\ World|Female) = \frac{0.05}{0.54} = 0.093$$
$$P(West\ World) = 0.25$$
$$0.093 \neq 0.25$$

由此可得，女性观众的概率和喜欢西部世界的概率是互相影响的(当然实际情况是几乎不可能相等，只会非常接近)

或者我们可以用这个公式:$If\ independent: P(A \cap B) = P(A) \times P(B), 0.05 \neq 0.54 \times 0{,}14$

如果想要求期望其实也是比较简单的，我们只需要跟以前一样，把每个的结果的值乘上他的概率再相加就可以

**期望&方差**

对于期望来说，其实跟以前的也差不多：就是每个可能发生的事件的随机变量的值*对应的概率，再相加就行

这是一个对于随机变量 X 的期望值:

$$E[X] = \int\int x f_{xy}(x,y) dy\, dx$$

可以把 x 提到外面一层

$$E[X] = \int x \int f_{xy}(x,y) dy\, dx$$

由于$\int f_{xy}(x,y)dy$其实就是边缘概率-对于 x=?所发生的概率的和

$$E[X] = \int x f_X(x)\, dx$$

就像之前的期望一样，我们也可以在联合概率的期望中放入函数：

$$E[g(X,Y)] = \sum_i \sum_j g(a_i, b_j) f_{XY}(a_i, b_j)$$

$$E[rX + sY + t] = rE[X] + sE[Y] + t$$

$$\begin{aligned} &Var[X+Y] = E[(X+Y)^2] - E[X+Y] \\ &= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \\ &= Var[X] + Var[Y] + 2(E[XY] - E[X]E[Y]) \end{aligned}$$

最后得到的这个绿色的部分，我们称作 covariance. 在统计学中，协方差（Covariance）是衡量两个随机变量联合变化趋势的一个指标. 如果两个变量的协方差为正，那么它们倾向于一起增加或减少；如果协方差为负，一个变量增加时另一个倾向于减少.

如果两变量无关，那么协方差一定为 0，但是如果协方差为 0，代表的是两变量有可能是无关，但不是绝对

换一种写法就是：

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

这种写法也是比较好理解的，因为我们就想知道两个变量的相关性，所以我们就把每个变量 X 减去所有 X 的样本的均值，乘上每个变量 Y 减去所有 Y 样本的均值，最后得到的均值就是协方差

## Correlation

由于$Var(X) = E[(X - E[X])^2]$，所以如果我们求$Cov(X,X)$，其实就是等于 X 的方差. 并且，有一个性质是：

$$Cov(X,Y)^2 \leq Cov(X,X) * Cov(Y,Y)$$

所以，我们这样写，就能得到 Correlation - 相关系数. 其中$\sigma$是标准差，就是方差开根

$$\frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \in [-1,1]$$

如果这个系数贴近 0，说明这两个变量不相关，否则离 0 越远越相关

### Bivariate Normal Distribution

## Intro To Statistics

首先，我们需要知道 iid 是什么：独立且来自同一个分布的随机变量$X_1, X_2$ … 虽然我们不知道他们来自什么分布，但我们知道的是他们都是独立的．接下来都是基于这个思想

实现-Realizations 是指随机变量的数值，如果观测随机变量$X$得到了一个具体数值$x$,那这个$x$就是$X$的实现

统计就是指对随机样本$(X_1, X_2, \dots, X_n)$的函数$T(X_1, X_2, \dots, X_n)$，我们虽然已经知道所有数据但还是需要统计来中随机样本进行计算是因为很多时候数据量太过庞大，我们使用抽样调查就可以预估总体.

如我们想算样本均值就是

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

而对于 n 个样本的方差则是(我们从数据中抽出一个样本-包含 n 个数据，然后用这些数据计算方差)

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

## Sample Distribution of Mean

对于一个采样结果:$(X_1, X_2, X_3, \dots X_n)$，它的$E[X_i] = \mu, Var(X_i) = \sigma^2$，则我们会知道对于样本均值的抽样分布$\bar{X}_n$(其实就是抽取 n 个样本，得到一个均值，做很多次，形成一个分布):

$$E[\bar{X}_n] = \mu$$
$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

并且还有一个特殊性质是**中心极限定理**：如果我们的样本足够大，重复很多遍后，这些均值组成的分布最终会趋近于正态分布，且均值是$\mu$，方差是$\frac{\sigma^2}{n}$

## Chi-Square Distribution

卡方分布与之前的不太一样，它有一个 degrees of freedom $k$，对于 k 个采样，我们把每个实现平方再相加

$$X_i \sim N(0,1), \qquad Y = \sum_{i=1}^{k} X_i^2, \quad Y \sim \chi^2(k)$$

$X_i \sim N(0,1)$的意思就是这个随机变量服从均值为 0，方差为 1 的正太分布

其实可以把之前算方差的公式$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$连起来:

$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$

上面的意思就是说样本方差$S_n^2$乘以$n-1$再除以总体方差$\sigma^2$，得到的统计量服从$k = n-1$卡方分布(当然单个统计量无法直接看出是否服从某种分布，还需要进行假设检验)

## Estimation

**教的答辩**

对于连续的变量，我们是比较好近似的：之前学过这么多的分布，总有一个适合你．但对于我们采出来的样，具体是用什么办法去近似成一个分部呢?

可能首先想到的是直接放进 R 或者其他程序中直接画图．但这样可能会给我们误导．更好的选择是先选择一个我们认为的分布(假设这个随机变量是遵循 xx 分布)．对于这个行为，我们可以成为 modeling – 建模.

然后对于选出来的分布，我们再预估它的参数(如正太分布就是有两个参数$\mu, \sigma^2$) $\theta$.

最终，我们的问题就变成了对于一堆采样$x_1, x_2, x_3, \ldots$他们遵从某个分布$f(\theta)$，这个 $\theta$是什么我们预估出来的大多数时候都与实际的不符，所以我们自己预测的就叫做$\hat{\theta} = T(x_1, x_2, \ldots)$

既然都开始预测了，那么我们的采样也会对我们预测出来的$\hat{\theta}$有所影响，所以我们认为这个$\theta$是个常量，而我们预测的$\hat{\theta}$也是个随机变量，如果我们想固定下来，就求$E[\hat{\theta}]$，如果最终的$E[\hat{\theta}] = \theta$，那么我们就称我们的$\hat{\theta}$为 unbiased estimator，反过来，我们称偏差$bias(\hat{\theta}) = E[\hat{\theta}] - \theta$

当然，对于一系列的采样，有可能我们通过两种算法得到了两个$\hat{\theta}$，他们都有相同的期望，我们如果想知道哪个$\hat{\theta}$预测的更好的话，可以看他们的方差，如果他们的 variance 越小，说明这个预测越有效．所以在我们找$\hat{\theta}$的时候，我们不仅需要最少的 bias，还需要最小的方差．

对于一般的分布来说，我们就正常计算就行：如对于正太分布中的$\mu$，我们预估量$\hat{\theta}$的计算方法就是$\frac{1}{n}\sum_{i=1}^{n} X_i$，方差的计算方法就是$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$，计算出来的就是 unbiased estimator.

例：比如我们要扔一个硬币 n 遍，得到随机变量$f_1, f_2, f_3, \dots f_n \sim_{iid} B(p)$ 每个随机变量都是独立的一次硬币结果，并且服从概率为$p$的伯努利分布．我们现在的目标就是找到这个 p，或者说$\hat{\theta} = T(x_1, x_2, \dots)$.

如，我们说这个函数$T = \frac{\min(f_1, f_2, \dots f_n) + \max(f_1, f_2, \dots f_n)}{2}$这个是计算 uniform Distribution 的公式，但在这里，我们很大概率也能得到$\frac{1}{2}$，这也是个 unbiased estimator.

或者，我们也可以说函数$T = \frac{1}{n}\sum f_i$这也是一个无偏估计.

后者看起来更靠谱一些，但我们得证明他为什么更好：

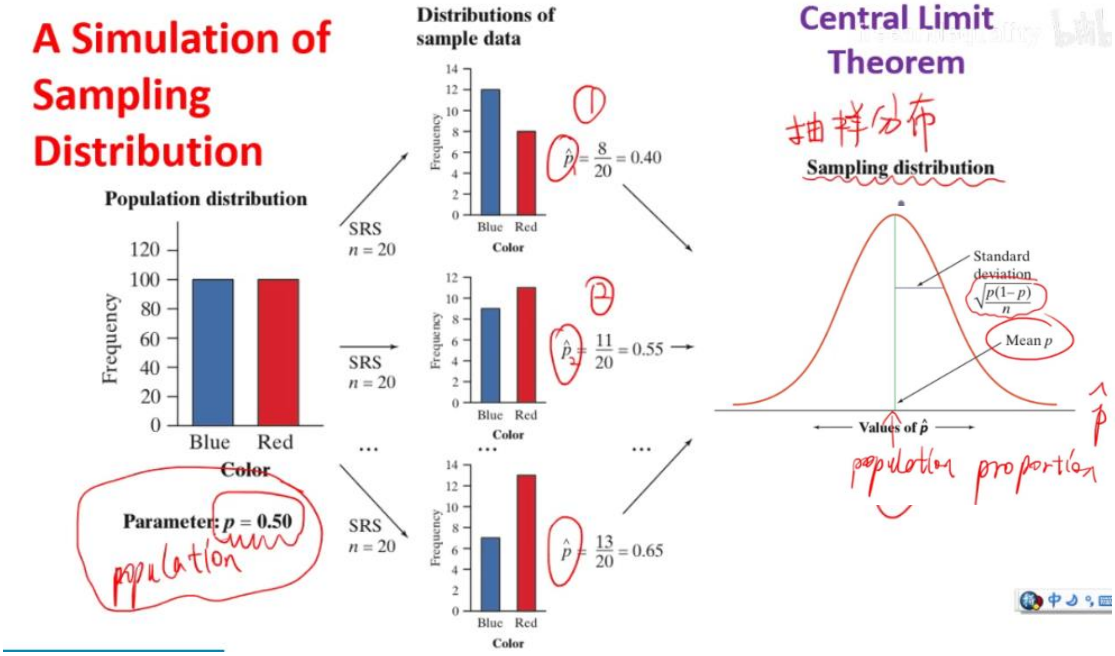
上面的计算公式可以写成这个:$\hat{p} = \hat{\theta} = avg(f_1, \dots f_n)$，$E[\hat{p}] = p$，因为这是伯努利分布，所以$E[f_i] = p$根据中心极限定理，样本均值就会等于正态分布的均值$\overline{f_n} = \hat{p}$

$$E\left[\overline{f_n}\right] = E[\hat{p}] = E[f_i] = p$$

这样，我们就证明了这是个真正的无偏估计，就算$p \neq 0.5$，我们也可以得到真正的结果，不像上一个如果样本足够多，就算$p \neq 0.5$，得到的$\hat{p}$永远都是0.5

**正片**

对于一个整体，我们经常通过抽样来试图参透全局．如一个袋子里红蓝球各有 100 个，我们让每个人抽 20 个，一直抽很多很多遍，每次抽出来的都是一个样本，样本就能得到一个大概的红蓝球比例(如我们求蓝球的比例就是$\hat{p} = \frac{x}{20}$)，如果我们每次抽的数量够大，这个$\hat{p}$就会越来越接近 0.5．我们可以画出一个图来表示这些样本，而随着我们这个样本数量的变多，这个图会越来越像正态分布




这个正太分布的特性就是：标准差 Standard deviation 是$\sqrt{\frac{p(1-p)}{n}}$，均值是$p$，请注意，虽然它的横轴是$\hat{p}$-Sample proportion，但画出来的正态分布的均值会是整体的比例，并且这里面的$p$都不戴帽子，也就是说不是由我们样本决定的，而是采样次数多，样本足够大后(但不能超过总体的$\frac{1}{10}$)，自动就会形成的真理

上面这种一次一次抽 Sample 来估计总体的叫点估计 - 这个点估计就是一个给我们提供了总体分布的参数(population parameter)的预估(estimate)的一个统计量(statistic)

*统计量就是任何我们通过 sample 算出来的东西, population parameter 就是总体的 mean, Standard deviation...*

虽然抽出来的 sample 得到的预估并不是一直等于总体的参数，我们仍然说这个是一个无偏估计，因为它的期望是等于总体的参数(均值).

但问题是虽然他是无偏，但是每次预测出来的确实不一样，我们还想知道这个预测出来的范围是什么，这就是置信区间

## Confidence Interval(Z 分布)

首先，我们根据中心极限定理，我们会知道当样本足够大的时候**样本均值**就会服从

$$\bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

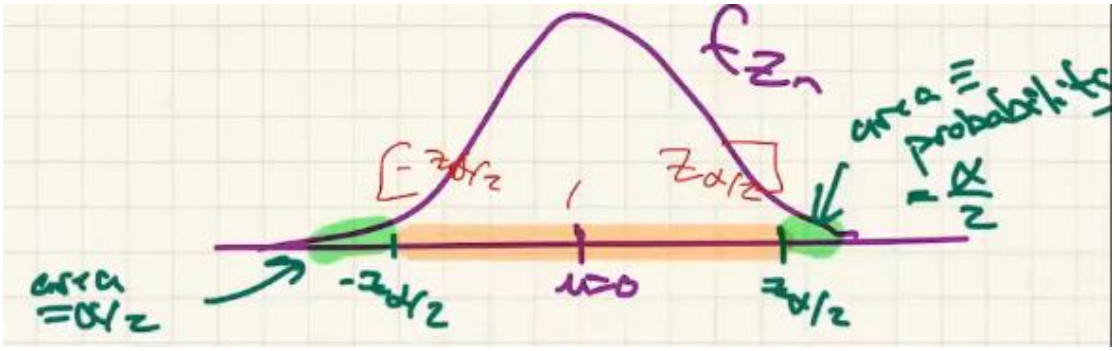
其中的$\sigma$是总体的标准差，$\bar{X}$是样本均值，$\mu$是总体均值，$n$是样本大小

我们就称作是 Z 分布，大致的算法就是我们把我们得到的数据标准化得到 Z 值，再根据 Z 表或计算器来算出样本得出的 Z 出现在这个 01 正态分布的百分之多少

**教的答辩**

当然，并不是所有的估计都是完美的无偏估计，所以我们需要某种办法去量化我们预测的参数的准确性：我们已经预测出了$\hat{\theta} = T(x_1, x_2, \dots)$，现在我们要加两个额外的函数，他们跟$T$差不多：$L_n = func(x_1, x_2, \dots)$，$R_n = func(x_1, x_2, \dots)$ 这两个函数分别也会得到两个随机变量. 我们最终就是要求$\Pr(L_n \le \theta \le R_n) = 1 - \alpha$ (这里的$\theta$已经定下来了)而$L_n \le \hat{\theta} \le R_n$这个区间就是$100(1-\alpha)\%$置信区间 但由于是对称的，所以我们只需要两个函数的其中一个就行.

例：对于一个已知方差的正太分布$x_1 \dots x_n \sim_{iid} N(\mu, \sigma^2)$，但我们不知道它的$\mu$我们就可以用置信区间去估计它的均值. 首先需要先标准化这个样本$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ 得到的$Z_n \sim N(0,1)$会服从正



态分布．接下来我们就想去求$\Pr\left(-Z_{\frac{\alpha}{2}} < Z_n \leq Z_{\frac{\alpha}{2}}\right) = 1-\alpha$

回到之前的问题，我们就要预测$\mu$，其实就是在求

$$\Pr\left(-Z_{\frac{\alpha}{2}} < \mu \leq Z_{\frac{\alpha}{2}}\right) = 1-\alpha \rightarrow \Pr\left(\bar{X} - Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} < \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

其中$\sigma$已经知道了，$Z_{\frac{\alpha}{2}}$也知道了(如 95%置信区间的 Z 就是 1.96), n 取决于我们抽多少个样本．真正的 random Variable 其实是被减的$\bar{X}$

根据 CTL，$Var[\bar{X}] = \frac{\sigma^2}{n}, std\ Var[\bar{X}] = \frac{\sigma}{\sqrt{n}}$

我们把上面的式子前半段进行一些变换，就能得到$\Pr\left(\bar{X} - \mu \leq Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) \rightarrow \Pr\left(\frac{(\bar{X}-\mu)}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\frac{\alpha}{2}}\right)$，

经过一番操作，就可以得到$\Pr\left(\frac{(\bar{X}-\mu)}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\frac{\alpha}{2}}\right) \rightarrow \Pr\left(Z_n \leq Z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$

说了这么多，其实不管有没有听懂，带公式总没错：

对于服从正态分布$X_i \sim N(\mu, \sigma^2)$的 Mean 的置信区间(已知 Variance)，我们可以这么算：首先算出 z-score 标准化样本，$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{Var[\bar{X}]}} = \frac{\bar{X}_n - \mu}{\sqrt{\frac{Var[X_i]}{n}}}$ 得到的 Z-score 就是在$N(0,1)$上的一个点．如我们想算$\Pr\ (\bar{X}_n \geq x)$，就$Z_n = \frac{x-\mu}{\sqrt{Var[\bar{X}]}}$，然后算$normCdf(lower = Z_n, upper = \infty, \mu = 0, \sigma = 1)$

**正片**

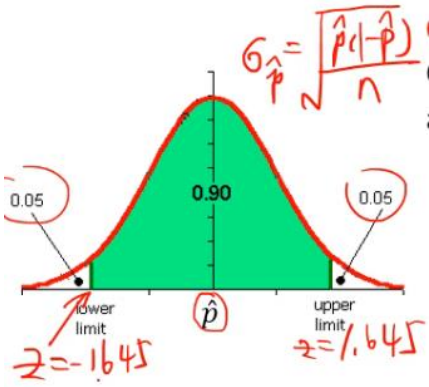
【AP 统计, AP Statistics】 https://www.bilibili.com/video/BV1WJ411Y76N/?p=19

如果我们想预测这个$\hat{p}$，光用一个 Sample 算出一个值肯定是不行的，它一直在变，因为是 random Variable，所以我们需要一个范围去包住它．这就是置信区间．如果我们说 90%的置信水平，就说明我们有 90%的信心保证预测出来的$\hat{p}$在这个区间里

比如一个框子里有红绿苹果，抽一个 Sample 251 个苹果，其中 107 是红的，我们想算 90%的置信水平的红苹果的比例：

$$Sample\ Proportion\ \hat{p} = \frac{107}{251} = 0.426$$

就像之前的红蓝球一样，这个也是伯努利分布，所以画出来的 Sampling Distribution 的标准差是$\sqrt{\frac{p(1-p)}{n}}$，均值是$p$，虽然这两个参数是由总体$p$决定的，但如果我们的样本足够多，就可以把$\hat{p}$当成样本来用.




最终画的图在右侧. 这里的 z score 其实是 table 给的，意思是对于一个$N(0,1)$分布, 95%的面积都在 1.645 面前，所以剩下的就是 0.05，两边一加就是 0.1，剩下的中间面积就是 0.9，因为这是 Standard Normal Distribution，所以还要乘上原本的标准差给变回去.

这样，我们就得到$Interval = statistic \pm critical\ value * standard\ Error = \hat{p} \pm z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

叫$standard\ Error$而不是标准差是因为我们把里面的$p$换成了$\hat{p}$，可能会造成偏差，而$z * standard\ Error$叫 Margin Error

但由于我们是通过 sample 计算出来的置信区间，所以每次出来的置信区间也不一样，有可能就某个置信区间没有框柱 population proportion，而我们的 confidence level 就是说 100 个 sample 求出来的 100 个置信区间，会有 90 个能框柱 population proportion

## T 分布

【十分钟理解 t 分布和 t 检验】 https://www.bilibili.com/video/BV1Qv411W7GQ/

在上面，我们都是用正态分布来进行估计，但是当样本不是足够大的时候，他也就不是正态分布，或者说，我们根本就得不到总体的$\sigma$. 我们就需要用 t 分布去估计. T 分布在样本小的时候会比较好用，因为他会给我们更高的容错率

所以，不得已，我们只能用样本标准差$S$来替换总体的标准差$\sigma$

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1), \qquad S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$Interval = \hat{p} \pm t * \frac{S}{\sqrt{n}}$$

这样求出的 t 他会服从 t 分布，这个 t 分布有一个参数是 dof，由样本-1 得来

当这个自由度越来越大(样本越来越大)，这个 t 分布也会越来越接近标准正态分布

请注意，这个并不是 t 检验，本质上还是 z 检验，只是因为我们替换了整体$\sigma$导致 z score 变成了服从 t 分布而不是标准正态分布的统计量

**总结:**

如果想计算**比例**的置信区间:

$$Interval = \hat{p} \pm z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

如果不知道总体的方差，我们一样可以用上面的公式估计比例

如果想计算样本均值的置信区间(已知总体$\sigma$):

$$Interval = \bar{X}_n \pm z * \frac{\sigma}{\sqrt{n}}$$

如果不知道总体的方差，则需要把总体方差$\sigma$替换成样本方差$S_n$

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1), \qquad S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad Interval = \hat{p} \pm t * \frac{S}{\sqrt{n}}$$

如果是双样本(独立样本):

$$Interval = (\bar{X}_1 - \bar{X}_2) \pm ME = (\bar{X}_1 - \bar{X}_2) \pm z * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= (\bar{X}_1 - \bar{X}_2) \pm t * \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \qquad t = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

如果想把样本均值转化成 Z - score(标准正态分布上的值) $Var[X_i]$是采样的方差

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\frac{Var[X_i]}{n}}}$$

## 假设检验

在之前的置信区间中，我们知道了如何求一个区间，这个区间会包含95%的 Sample mean，但实际应用是什么?

我们可以用在假设检验上面：假设检验可以告诉我们一个事情是否足够可信

比如我们有一个理论$H$，我们想要证明这个$H$对不对，我们可以:

- 先写出原假设$H_0$，就是原本我们想证明的假设
- 写出备择假设$H_a$，就是与我们原本的假设相反的假设
- 接下来采样，算出我们的均值，如果均值跟假设相差不大，则我们相信原假设，否则相信备择假设，写成公式就是若$(\bar{X} - \mu_0) < c$时接受，否则拒绝．也就是$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{c}{\sigma/\sqrt{n}}$时拒绝．后面的这个公式的前半部分是把我们的样本转换成 z score，后半部分是把我们的要求转化成 z score，也就是如果样本出现的概率比我们的最低要求出现的概率还要小(z 越大概率越小)，则拒绝假设
- 根据我们的计算过程检查是否可能出现一类/二类错误

如一个果汁厂商声称饮料中有果汁 500ml.

$$H_0: \mu = 500, H_a: \mu \neq 500$$

我们采样了 100 个样本，其中平均体积是 490ml. 我们将 P 值设成5% - 其实就是置信区间为标准正态分布的95%的面积，对应的 Z-score 是 1.96

根据上面的公式，我们求$\frac{490-500}{50/\sqrt{100}} = 2 \geq 1.96$，所以我们拒绝原假设

【30 分钟拿下区间估计与假设检验】 https://www.bilibili.com/video/BV1DP411a732/

在检验的时候，我们肯定会犯错：

Type 1 Error：当$H_0$正确的时候，我们还是拒绝了$H_0$，拒绝的概率为$\alpha$，在显著性测验的时候得出的 Z score 概率小于$\alpha$都会被拒绝. (例如我们抽的样本刚好都是不达标的样本)

Type 2 Error：当$H_0$错误的时候，没有拒绝$H_0$. (如我们抽到的刚好是达标的样本，剩下全是不达标的)

对于一个固定样本数量，这两个错误的概率不能同时降低. 如果我们想降低错误 1 的概率，就需要增加接受区间，如把置信区间从 90%增加到 95%

| 检验参数 | 情形 | 假设 | | 检验统计量 | $H_0$ 为真时检验统计量的分布 | 拒绝域 |
|---|---|---|---|---|---|---|
| | | $H_0$ | $H_1$ | | | |
| $\mu$ | $\sigma^2$ 已知 | $\mu=\mu_0$<br>$\mu\leqslant\mu_0$<br>$\mu\geqslant\mu_0$ | $\mu\neq\mu_0$<br>$\mu>\mu_0$<br>$\mu<\mu_0$ | $U=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ | $N(0,1)$ | $\lvert U\rvert\geqslant u_{\frac{\alpha}{2}}$<br>$U\geqslant u_\alpha$<br>$U\leqslant -u_\alpha$ |
| | $\sigma^2$ 未知 | $\mu=\mu_0$<br>$\mu\leqslant\mu_0$<br>$\mu\geqslant\mu_0$ | $\mu\neq\mu_0$<br>$\mu>\mu_0$<br>$\mu<\mu_0$ | $T=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ | $t(n-1)$ | $\lvert T\rvert\geqslant t_{\frac{\alpha}{2}}(n-1)$<br>$T\geqslant t_\alpha(n-1)$<br>$T\leqslant -t_\alpha(n-1)$ |

## T 检验

T 检验分为三种

- 独立样本 t 检验，这种是为了检验两组不同的样本的数据之间的平均值是否存在差异，如两班成绩是否一样

$$t = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

得到的 t 值如果比 t 值表对应的值大的话，就说明两组存在显著差异

- 配对样本 t 检验，这种是为了检验同一组样本得到的两组数据是否存在差异，如一班学生在上完补习班后成绩有没有变

$$t = \frac{\bar{d}}{S_d/\sqrt{n}}$$

这里的$\bar{d}$是样本差值的均值$d_i = X_{1i} - X_{2i}, \bar{d} = \frac{1}{n}\sum d_i$

$S_d$是样本差值的均值的标准差$S_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}}$

同样，我们需要看表得到是否存在显著差异

- 单样本 t 检验，这种是为了求一组样本均值和已知总体均值有没有存在显著差异，如一班的身高和全国身高有没有差异

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

这些公式得到的 t，是 t 分布横坐标上的一个点(图中是$u_0$)．它表示的是对于我们得到的样本均值转化到 t 分布上的值．我们通过查表可以得到拒绝阈值$u_{1-\alpha}$，如果这个值$u_0$超过了$u_{1-\alpha}$，我们就拒绝．

而还有另一种方法拒绝就是看面积，如果$u_0$之后的面积大于$u_{1-\alpha}$之后的面积，就说明$u_0 < u_{1-\alpha}$，我们就接受原假设．这个面积就是样本发生以及比样本更极端的情况发生的概率，也叫 p-value