1 Learning Outcome

Measure of Spread:

- ☐ **绘制频率直方图，累积频率直方图** histogram& cumulative frequency graph
- ☐ **计算离散统计值的方差** Variance **和标准差** Standard Deviation

Probability:

- ☐ **一个事件概率的表示手段**
- ☐ **绘制** tree diagram
- ☐ **互斥事件** mutually exclusive events **与独立事件** independent events **的判断**
- ☐ **求算条件概率** Conditional Probability

Permutations & Combinations:

- ☐ **阶乘** factorial **的表示方式和计算**
- ☐ **排列** Permutation & Arrangements **的计算公式和应用**
- ☐ **组合** Combination **的计算公式以及应用**
- ☐ **插空法的使用**

Probability Distribution:

- ☐ **理解随机变量的含义，离散的和连续的** Discrete and Continuous Random Variable
- ☐ **计算随机变量的概率分布表** Probability Distribution
- ☐ **根据概率分布表，求算随机变量的期望** Expectation **和方差** Variance

Binomial Distribution:

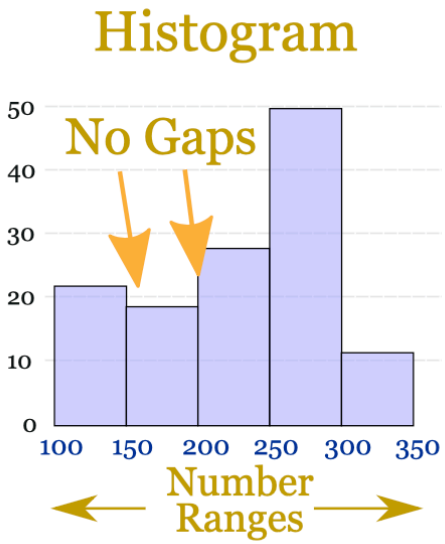
- ☐ 理解 0-1 分布实验
- ☐ 理解二项分布 Binomial Distribution 实验，掌握表达方式，求算特定次数的概率，列出概率分布表
- ☐ 求算二项分布的期望和方差
- ☐ 理解几何分布 Geometric Distribution 实验，掌握表达方式，求算特定次数的概率，列出概率分布表
- ☐ 求算几何分布的期望

The Normal Distribution:

- ☐ 掌握正态分布表达方式，概率曲线形状，期望和标准差对概率曲线的影响，密度曲线包围的面积的意义
- ☐ 对正态分布做标准化处理 Standardize Normal Distribution
- ☐ 能够从正态分布表查值
- ☐ 利用标准正态分布解决相关的概率问题

2 Key Concept

1. 频率分布直方图 histogram 是将结果按照一定的 interval 统计频率绘制的图像。如下图所示：



2. 累计频率分布表是将所有小于给定值的频率相加获取的，较大数值的累计频率必定大于等于较小数值的累计频率

x	frequency	range	cumulative frequency
10-20	4	<20	4
20-30	9	<30	13
30-35	7	<40	20
35-50	3	<50	23

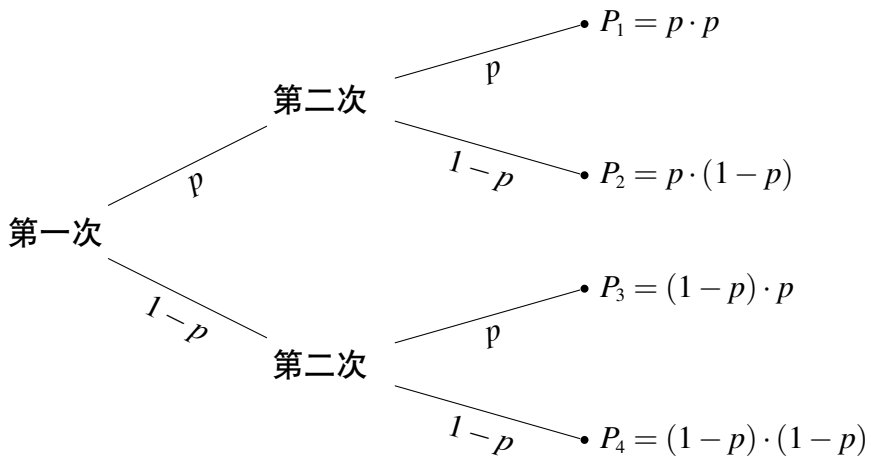
3. 对于带有频率分布的统计值，求算平均值 Mean的公式为：$\bar{x} = \frac{\sum x_i \cdot f_i}{\sum f_i}$。

4. 对于带有频率分布的统计值，求算方差 Variance的公式为：
$Var(X) = \frac{\sum (x_i - \bar{x})^2 \cdot f_i}{\sum f_i} = \frac{\sum x_i^2 \cdot f_i}{\sum f_i} - \bar{x}^2$。
标准差 Standard Deviation就是方差的正平方根。

5. 如果把一个事件记录为 A，另外一个时间记录为 B，那么 $P(A)$ 表示 A 事件发生的概率，$P(A')$ 表示A 的对立事件发生的概率 $P(B)$ 表示 B 事件发生的概率，$P(AB)$ 或者 $P(A \cap B)$

或者 $P(AandB)$ 表示 AB 同时发生的概率，而 $P(A|B)$ 则表示在B 事件已经发生的条件下，A 事件发生的概率。由于概率的定义，必定是介于 0-1 之间的一个数据。

6. tree diagram 是将所有可能的结果全部罗列出来的图表，并且在每一次的分支上可以记录发生的概率，沿着某一条路径分支一致走下去就是最终概率。并且所有分支的概率之和为 1。一般考试节点级数不超过 3 个。如下图：



7. 两个事件为互斥事件 Mutually Exclusive Events是指这两个事件不能同时发生。因此 $P(A \cap B) = 0$

8. 两个事件为独立事件 Independent Events 是指两个事件没有任何联系。一个事件的发生不会影响第二个事件的发生与否。因此 $P(A \cap B) = P(A) \times P(B)$，或者说 $P(A|B) = P(A)$，$P(B|A) = P(B)$

9. 条件概率Conditional Probability就是指当某一个事件发生后，另一个事件发生的概率。有非常著名的贝叶斯公式。$P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ 因此 $P(A|B) \geq P(A)$

10. 阶乘 factorial 是一种正整数连乘的缩写手段，$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$。其中通过阶乘的特殊性质，定义了 $0! = 1$

11. 排列是指从 n 个物体当中，取出 r 个物体，并且按照特定的顺序将这个 r 个物体进行排列，记录有多少种不同的排列顺序。比如学生站队的问题。那我们直接通过 Permutation 的求算。${}^nP_r = \dfrac{n!}{(n-r)!}$。核心关键词是Arrange，align

12. 组合是指从 n 个物体当中，取出 r 个物体，而无需考虑这 r 个物体的先后顺序，记录所有可行的抽选方案数目。可以直接通过 Combination 的公式进行求算 ${}^nC_r = \dfrac{{}^nP_r}{r!} = \dfrac{n!}{(n-r)! \cdot r!}$，可以认为是对应 Permutation，除以 $r!$ 种重复顺序得来的。另外一种标记手段是 $\binom{n}{r}$，在二项分布和二项展开式当中这种形式居多。

13. 插空法是常用的解题方法，通常用于多个物体必须要紧密相邻，或者多个物体不能相邻的情况。其思路是：

    a. 将没有做其他要求的剩余物体进行 Permutation；

    b. 数出有多少个空隙，记为 n；

    c-1. 如果要求是 r 个物体相邻，则将这 r 个物体进行 Permutaion，再任选空隙当中的一个进行插入 $r! \cdot \binom{n}{1}$；

    c-2. 如果要求是 r 个物体不能相邻，则从空隙当中选择出 r 个空，再有顺序地将物体插入至空隙中；因此结果为 $\binom{n}{r} \cdot r!$ 或者直接用 ${}^{n}P_{r}$.

    d. 最后利用乘法法则求算最终结果

14. 随机变量 Random Variable，一般用大写字母 X 表示，代表着某一个试验最终所有可能的结果。这里由于并没有真正去做这次实验，因此我们是从假设预估的角度上进行探究的。而离散的随机变量指 X 的取值为离散数值，一般为整数。连续随机变量则是指 X 能取到所有实数值

15. 离散随机变量的概率分布表就是如下图所示的一张表：

X	x_1	x_2	x_3	x_4
$P(X)$	p_1	p_2	p_3	p_4

    其中 x_i 是随机变量 X 的取值，下方则表示对应的概率。有一个很重要的性质经常会用于解题当中，由于这张表是所有可能取值，因此把所有概率加起来最后结果等于 1。在考试当中基本上不会给表格，而是给出概率随取值的函数关系 $P(x=k)=f(k)$，要能够理解这种函数关系，并且转化到这张表

16. 从概率分布表，可以求算该随机变量的期望和方差，需要牢记公式：

    期望 Expectation $E(X)=\sum x_i \cdot P(x_i)$

    方差 Variance $Var(X)=\sum (x_i - E(X))^2 \cdot P(x_i)$

17. 0-1 分布指随机试验的结果只有两种的分布情况，可以用 0 代表失败，用 1 代表成功。最为经典的例子为抛硬币。0-1 分布的表格为：

X	0	1
$P(X)$	$1-p$	p

18. 二项分布实验 Binomial Distribution，则是将 0-1 分布实验重复执行 n 次。用随机变量 X 来表示在这 n 次实验当中成功的次数。表达方式为为 $X \sim B(n,p)$。其中具有两个参数 parameter，实验重复次数 n 以及单次成功的概率 p。由于每次实验之间都是互相独立的，因此如果成功次数为 k 的话，概率为 $P(X=k)=\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}$。

X	0	1	…	k	…	n
$P(X)$	$(1-p)^n$	$\binom{n}{1}p^1(1-p)^{n-1}$	…	$\binom{n}{k}p^k(1-p)^{n-k}$	…	p^n

19. 因此使用该概率分布表，可以求算二项分布的期望和方差，分别为 $E(X)=np$，$Var(X)=np(1-p)$。这两个公式无需证明，只需要运用即可。

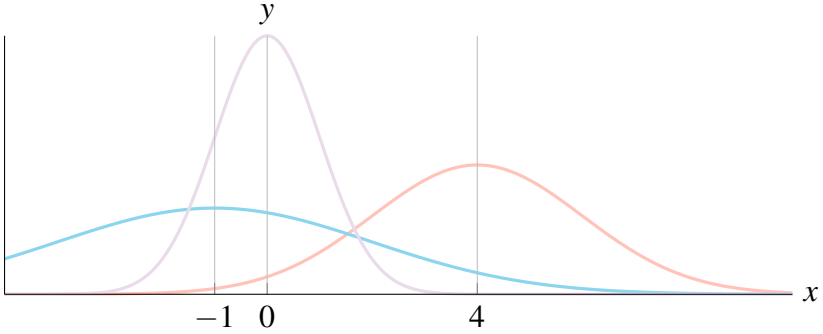
20. 几何分布 Geometric Distribution 也是从 0-1 分布演变而来的，但是和二项分布不一样的是，在几何分布实验中，并不是指定做多少次实验，而是要一直重复做 0-1 实验，直到成功。再用随机变量 X 表示总共需要的次数。计作 $X \sim Geo(p)$。其中只有一个参数，成功的概率 p。比如 $X=3$ 意味着前两次失败，第三次成功。因此求算几何分布的概率分布表则如下表所示

X	1	2	…	k	…
$P(X)$	p	$(1-p)p$	…	$(1-p)^{k-1}\cdot p$	…

几何分布和二项分布最不一样的地方在于这个概率分布表从 1 开始，并且无穷无尽。而且不管 X 取值为多少，只能是在最后一次成功

21. 仅考察几何分布的期望，求算公式为 $E(X)=\dfrac{1}{p}$

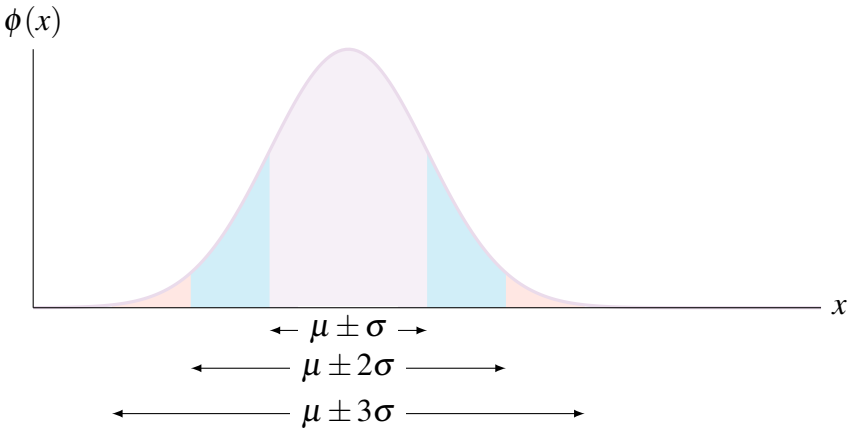
22. 正态分布 Normal Distribution 是最为重要的连续分布，也是日常生活当中非常常见的分布情况。不需要掌握正态分布的概率密度函数，仅需要记住正态分布的图像如下图所示：



23. 对于服从正态分布的随机变量，我们的标记手段为，$X \sim N(\mu, \sigma^2)$。其中，μ 为正态分布的期望，σ 代表标准差 standard deviation。从上图当中，可以看出任意正态分布的曲线都是类似的，但是 μ 会使分布图像水平移动，σ 则会使图像发生竖直方向的拉伸变形。并且特征是，当 σ 越大的时候，分布更加的分散，因此会压扁函数图像。

    上图的三个正态分布，分别为 $N(0,1^2)$，$N(4,2^2)$，$N(-1,3^2)$。尝试区分对应关系

24. 对于正态分布的图像，最重要的是图像下方包围的面积

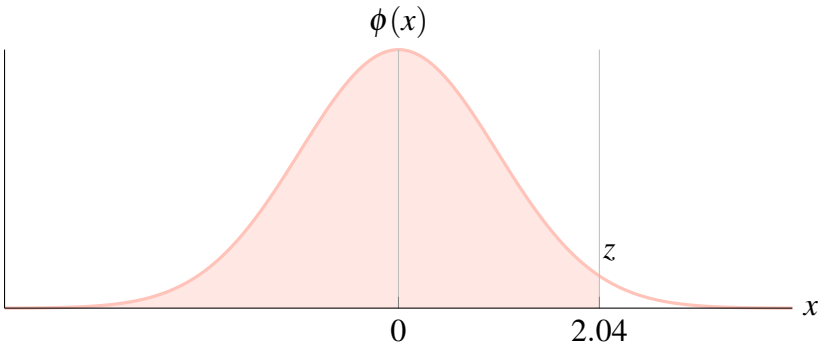


    上图所示，阴影部分面积分别表示随机变量的取值在偏离期望 1 个标准差，2 个标准差，以及 3 个标准差的的概率大小。$P(\mu-\sigma<X<\mu+\sigma)=68.3\%$，$P(\mu-2\sigma<X<\mu+2\sigma)=95.4\%$，$P(\mu-3\sigma<X<\mu+3\sigma)=99.7\%$ 这个关系对于任意的正态分布总是成立的。

25. 既然所有的正态分布都可以通过拉伸变形到同样的图像，我们选择 $N(0,1^2)$ 的正态分布作为这个基本图像，称之为标准正态分布 Standard Normal Distribution，那么其他任意分布都可以通过移动和变形回到标准正态分布的曲线上。这个过程叫做标准化 Standardization。

    假设 $X \sim N(\mu, \sigma^2)$，那么引入新的随机变量 Z，使 $Z=\dfrac{X-\mu}{\sigma}$，这个随机变量 Z 会服从标准正态分布。$Z \sim N(0,1^2)$。因此 $P(X>k)$ 就完全等价于 $P(Z>\dfrac{k-\mu}{\sigma})$，到标准正态分布表里去找值就好了。

26. 标准正态分布表是如下图的一张表，是求算 $N(0,1^2)$ 分布中，随机变量小于 z 值的概率。$\Phi(z)=P(X<z)=\int_{-\infty}^{z}N(0,1^2)\mathrm{d}x=\int_{-\infty}^{z}\phi(x)\mathrm{d}x$:



	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
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

使用这张表的过程是比较简单的，比如查找 $\Phi(2.04)$ 直接先搜寻第一排竖排 2.0，再水平移动至第五个，选择 0.04 这一栏。找出交点数值为 0.9788。就表示在 $N(0,1^2)$ 分布中，随机变量小于 2.04 的概率为 0.9793。考试当中也可能反向查找，比如给定概率为 0.9788 寻找对应的 z 值。