假设即将举行总统选举，候选人有两位：A和B。假设这个国家每个人都投票，这里存在一个百分比。投给B的百分比是P，投给A的百分比为1-P，能得到的就是两个值

为了能够进行运算，我们定义投票给A记为0，投票给B为1。根据伯努利分布，这个分布的均值为P。

但是，我们不可能一个个问别人投票给谁。不过可以进行一项随机调查。从总体中进行抽样，然后根据样本情况估计P值。比如，随机调查100个人的样本。假设结果如下：

```
57个人选A

43个人选B
```

那么样本均值：

$$\bar{X} = \frac{57 \cdot 0 + 43 \cdot 1}{100} = 0.43$$

样本方差：

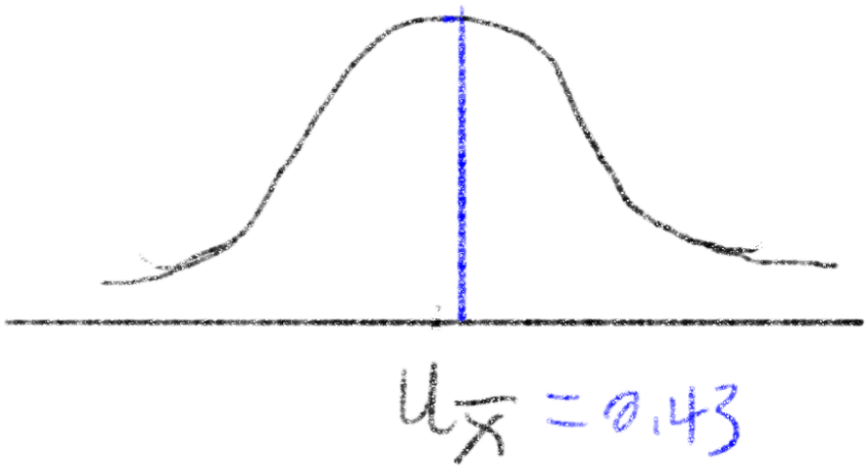
$$S^2 = \frac{57(0 - 0.43)^2 + 43(1 - 0.43)^2}{100 - 1} = 0.2475$$

标准差：

$$S = 0.497 \approx 0.5$$

这样，最好的估计值就算出来了，43%的人投给B，57%的人投给A。但是，这个样本有多好呢？我们确信95%的可能性会在0.43周围的区间内。

样本均值相当于来自样本均值的抽样分布。抽样分布的均值$\mu_{\bar{X}}$又等于总体均值：



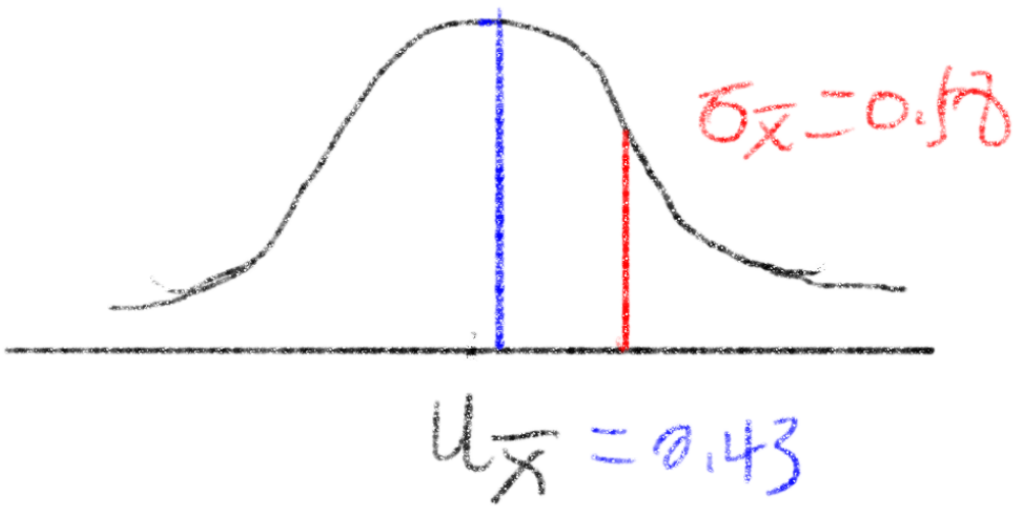
> 抽样也就是总体中抽样。所以总体的均值μ等于P。

再来看看这个分布的标准差。样本均值抽样分布的标准差等于总体标准差σ除以根号n：

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{100}} = \frac{\sigma}{10}$$

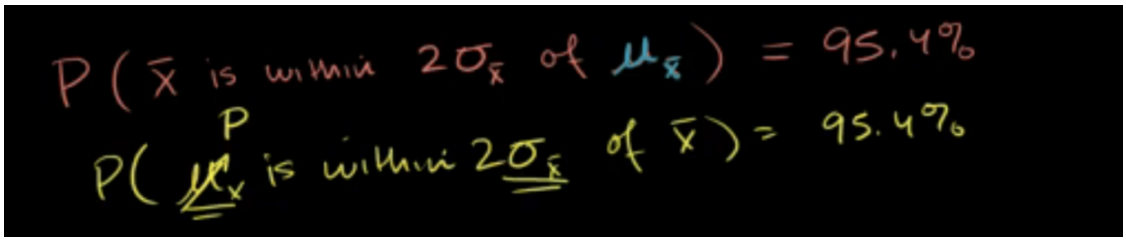
问题是，不知道σ是什么。我们只能用样本标准差S来估计总体的标准差σ。 样本标准差是对总体标准差的最好的估计。如上面案例中，总体标准差的估计为0.5，则获得抽样分布的标准差：

$$\sigma_{\bar{X}} = \frac{0.5}{10} = 0.05$$



然后，我们要找的的是样本均值周围的一个置信区间，使真正的总体均值有95%的可能落在此区间内。

真实的均值也就是样本均值抽样分布的均值。而样本均值落在抽样分布均值周围2个标准差的概率是95%；反之亦然，抽样分布均值在样本均值的2个标准差内的概率也是95%。抽样分布的均值等于原分布的均值，也就是P，所以P在样本均值的2个标准差内的概率也是95%



我们前面估得的抽样分布的标准差为0.05，也就是说P在样本均值周围0.10以内的概率为95%。

样本均值周围0.1也就是0.43 ± 0.1这个区间内，真实均值落在这个区间的概率为95%。这里得到了一个33%到53%的95%可信的置信区间。换句话说：

> 43%的人会投票给B
> 57%的人会投票给A

然后，人们给出误差范围。这里的误差范围是10%，也就是说，虽然43%的人投票给B，但是，它仍然有可能（$43\% + 10\%$）赢得大选。

> 误差范围是描述置信区间的另一种方法（margin of error is just another way of describing the confidence interval）。

如果想要更准确，需要增加样本。如果样本容量1000，抽样分布的标准差会减少，于是误差范围也会减少。

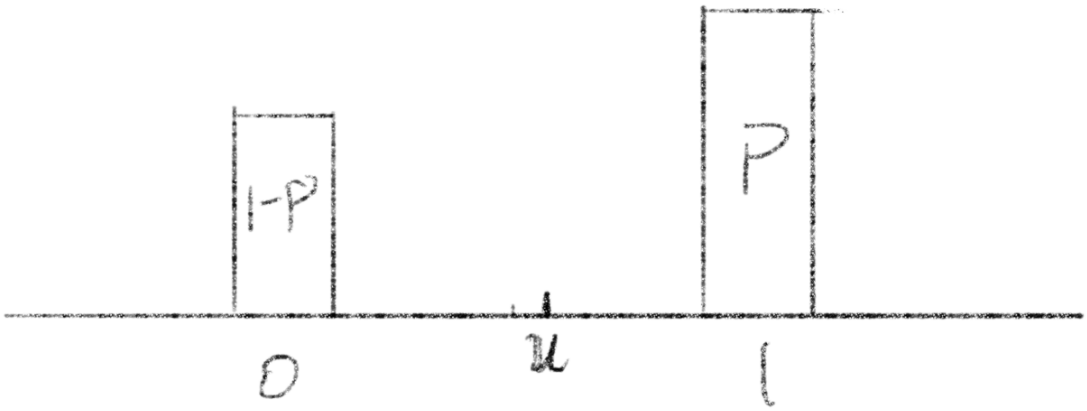
例题：某地方教学区获得一笔拨款，用于在教室安装4台一组的计算机。从6250个教师中随机抽取250人问计算机是否教室必备工具。这些人中142个认为计算机是必须的。

问题1： 计算一个99%的置信区间，其中教师认为计算机是必备的教学工具。

首先，我们定义随机变量：

$$X = \begin{cases} 0, \text{认为计算机不是必备工具} \\ 1, \text{认为计算机是必备工具} \end{cases}$$

绘图如下：



因为，我们不能调查每一个人，所以进行抽样，样本为250。

求得样本的均值：

$$\bar{X} = \frac{0 \times 108 + 1 \times 142}{250} = 0.568$$

样本的方差：

$$S^2 = \frac{108(0-0.568)^2 + 142(1-0.568)^2}{249} = 0.246$$

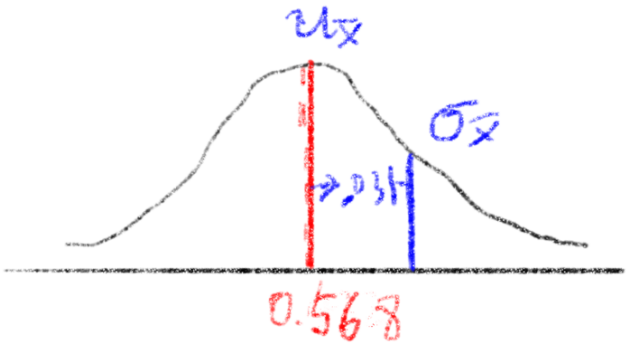
样本的标准差：

$$S = \sqrt{S^2} = 0.496$$

抽样是从样本均值抽样分布中进行的抽样。抽样分布大概如下：

> 抽样分布的均值等于原分布的均值，也就是P。

> 抽样分布的标准差等于原分布的标准差除以根号下样本容量n。但总体标准差是未知的，所以，我们使用它的最好的估计——样本标准差。之所以叫置信，是因为我们有信心，总体标准差落在这个区间内。
> $\sigma_{\bar{x}} = \frac{0.496}{\sqrt{250}} = 0.031$



现在，我们要的是一个99%的置信区间。可以这样考虑，随机从抽样分布中抽取一个样本均值，离均值多少个标准差范围内我们能99%相信抽取的样本值落在这个区间内。

要求出这个，我们需要借助z分数表（Z-Table）。求得99%，也就是找到0.995在z分数表上的位置。

> 因为正态分布的对称性，求99%的区间也就是求均值右侧为49.5%的概率。 又因为，z分数的值指的是小于该标准差的所有的概率，所以我们加上均值左侧的所有概率。因此，要求的就是49.5%+50%=99.5%在z分数表上的位置：

查表：

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981

结果为2.58个标准差。也就是误差范围为：$\pm(2.58 \times 0.031) = \pm 0.08$

最后，99%的置信区间为：

0.568 ± 0.08

也就是：0.488~0.648。

认为计算机是必备工具的人的概率在48.8%到64.8%之间的可信度为90%