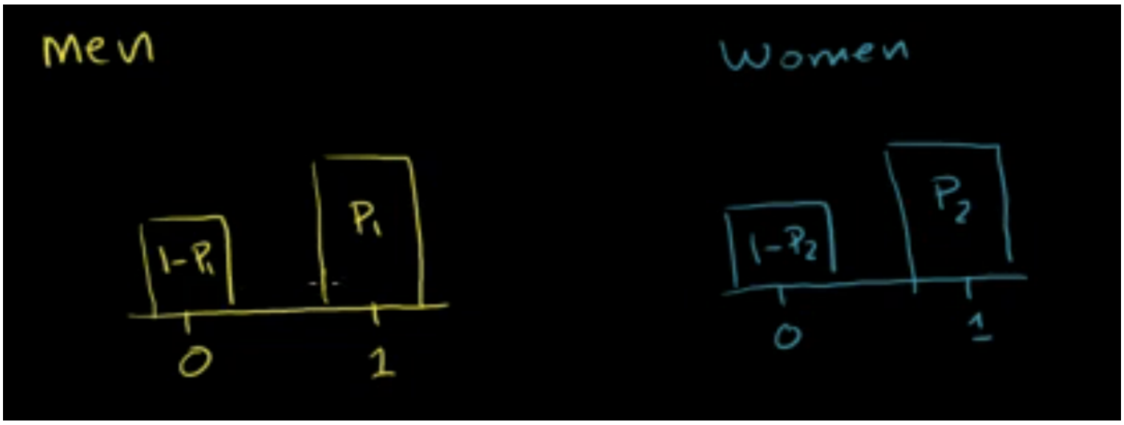
假设将有选举临近，我想知道男性和女性中，投给某候选人的占比是否有显著不同。看一下总体分布：



根据伯努利分布的性质，均值等于投给此候选人的占比值。

1. 总体比较

男性投票的均值与方差：

$$\mu_1 = P_1$$
$$\sigma_1^2 = P_1(1 - P_1)$$

女性投票的均值与方差：

$$\mu_2 = P_2$$
$$\sigma_2^2 = P_2(1 - P_2)$$

要求男性和女性投票之间是否有显著差别，也就是求P_1和P_2之间是否有显著差别: $P_1 - P_2 = ?$。参数之差任然是参数，我们不知道具体的值是什么，但我们希望得到一个95%的置信区间。

为此，我们调查了1000个投票的男性和1000个投票的女性。1000个男性中642选择1；1000个女性中591选择1。

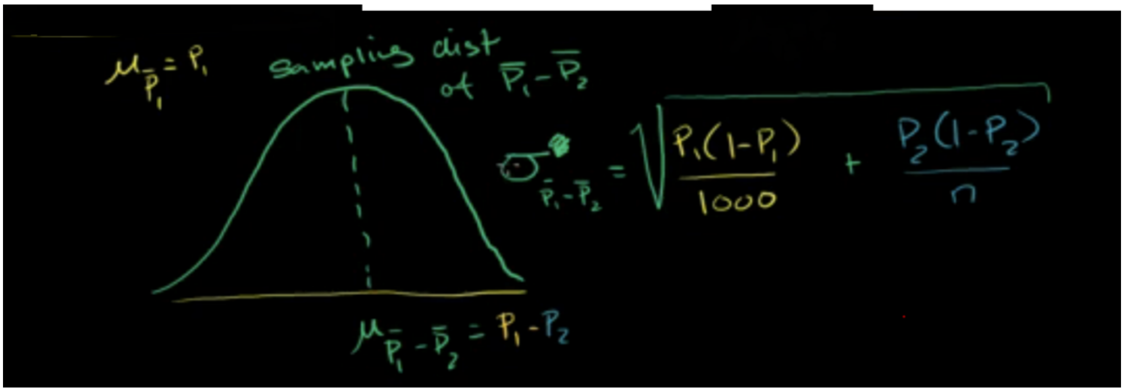
男性和女性的样本均值为：

$$\bar{P}_1 = 0.642$$
$$\bar{P}_2 = 0.591$$

男性和女性抽样分布的方差：

$$\sigma^2_{\bar{P}_1} = \frac{P_1(1-P_1)}{1000}$$
$$\sigma^2_{\bar{P}_2} = \frac{P_2(1-P_2)}{1000}$$

我们还要考虑两样本占比之差的抽样分布。样本占比可以认为是抽样分布的一个样本值。两个样本均值之差的所有可能性构成了均值之差的分布：



我们来看看具体的数值。样本之差的均值：

$$\bar{P}_1 - \bar{P}_2 = 0.642 - 0.591 = 0.051$$

我们希望有95%几率，实际均值$P_1 - P_2$落在这个样本差值0.051左右某个范围内。因为样本量很大，这个95%的可信范围通过查Z表中97.5%为1.96。这个范围的极限值就是1.96乘上分布的标准差。

抽样分布的标准差：

$$\sigma_{\bar{P}_1 - \bar{P}_2} = \sqrt{\frac{0.642(1-0.642)}{1000} + \frac{0.591(1-0.591)}{1000}} = 0.022$$

这个置信区间为：0.051 ± 0.043。也就是(0.008,0.094)之间。因此投给某一特定候选人的男女占比之差的95%的置信区间是0.8%到9.4%之间

2. 假设检验

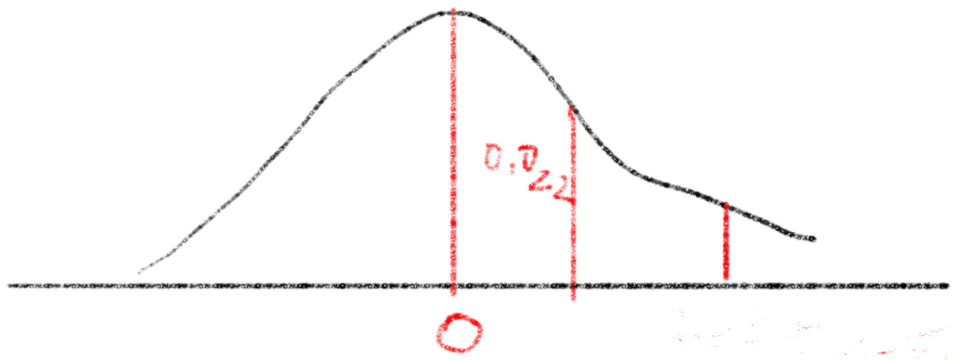
(1)首先进行假设：

$$H_0: P_1 - P_2 = 0 \qquad \text{no difference}$$
$$H_1: P_1 - P_2 \neq 0$$

(2) 求实际样本占比差值的概率

在零假设成立的前提下，求出实际样本占比差值的概率，如果该概率小于显著性水平5%。我们将距离零假设。

占比差值的分布如下：



求z分数：

$$Z = \frac{0.051 - 0}{0.022} = 2.34$$

0.051在均值0外2.34个标准差远，这个的概率小于显著性水平(5%),它比临界Z值的情况更极端，所以，我们拒绝零假设，更倾向于男女投票占比之间存在差异。