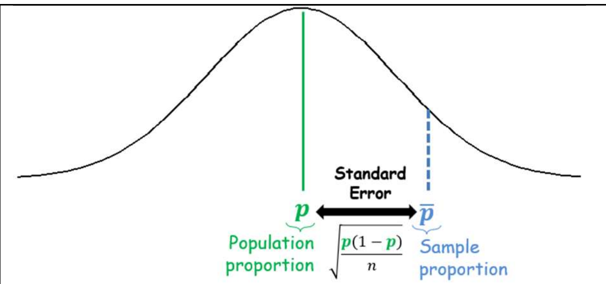
Deriving the Confidence Interval

Recall, the **sample proportion ($\overline{p}$)** follows a Normal Distribution.

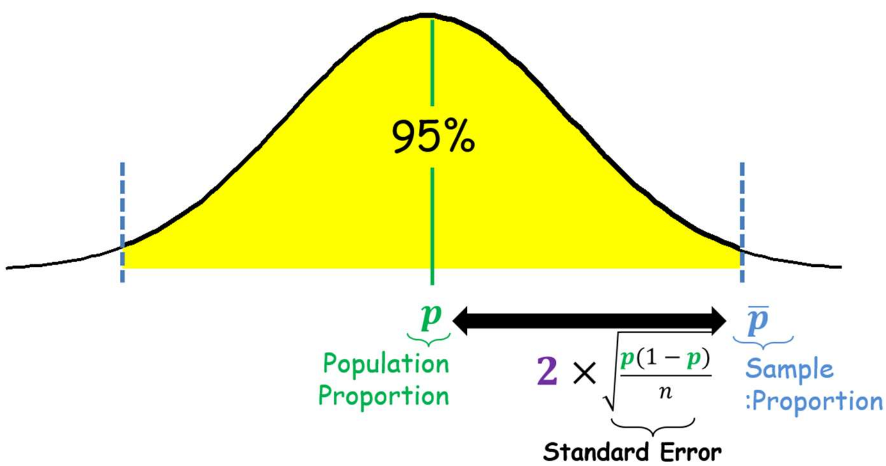
- The mean is given by the **population proportion (p)**
- The **standard deviation** is given by

$$SD[\overline{p}] = \sqrt{\frac{p(1-p)}{n}}$$

(This term is also called **Standard Error**)



According to the 95% rule for a Normal Distribution, **95% of data (e.g. sample proportions)** fall within **2 standard deviations (or standard errors)** of the **population proportion**.



If we assume our sample proportion falls in this 95% range,

then the difference between **the sample proportion ($\overline{p}$)** and its **population proportion (p)** is given by at most **2** x **Standard Error.**

$$|p - \overline{p}| \leq 2 \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error}$$

Then we can rewrite the expression in terms of p. We have the following expression.

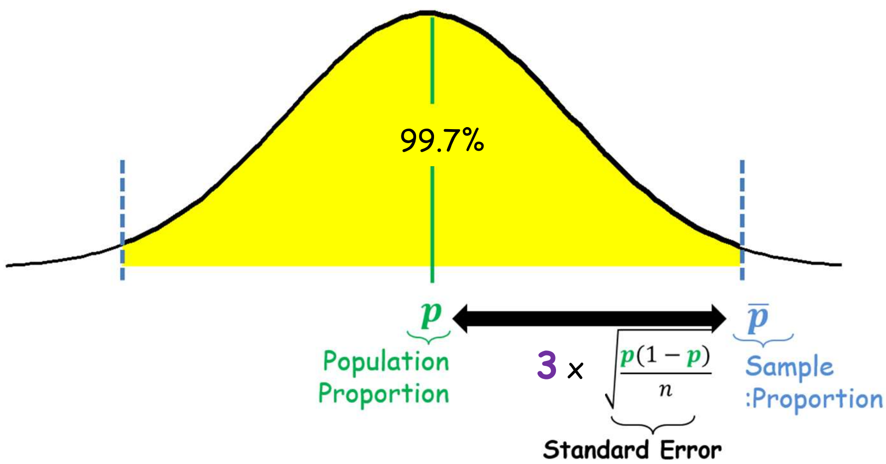
$$\Rightarrow \quad \overline{p} - 2 \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error} \leq p \leq \overline{p} + 2 \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error}$$

The expression provides an **interval** that captures the **population proportion** with a 95% probability.

Deriving the Confidence Interval

If you want to increase the probability that the interval captures the population proportions, we can use the 99.7% rule for Normal Distribution.

According to the 99.7% rule for a Normal Distribution, **99.7% of data (e.g. sample proportion)** fall within **3 standard deviations (or standard errors)** of the **population proportion**.



If we assume our sample proportion falls in this 99.7% range,

then the difference between **the sample proportion ($\overline{p}$)** and its **population proportion (p)** is given by at most **3** x **Standard Error.**

$$|p - \overline{p}| \leq 3 \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error}$$

Then we can rewrite the expression in terms of p. We have the following expression.

$$\Rightarrow \quad \overline{p} - 3 \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error} \leq p \leq \overline{p} + 3 \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error}$$

The expression provides an **interval** that captures the **population proportion** with a 99.7% probability.

Deriving the Confidence Interval

In general, if you want to construct an interval that captures the population proportions with a desired probability, you only need to multiply with the standard error with the right **constant**.

$$\bar{p} - \boldsymbol{constant} \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Standard\ Error} \leq p \leq \bar{p} + \boldsymbol{constant} \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Errpr}$$

The value of this constant depends on the desired probability level and is obtained from the **Normal distribution**. For this reason, it is called the **z-critical value**.

$$\bar{p} - \underbrace{z_c}_{critical\ value} \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Standard\ Error} \leq p \leq \bar{p} + \underbrace{z_c}_{critical\ value} \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error}$$

Therefore, the expression $\bar{p} \pm \underbrace{z_c}_{critical\ value} \times \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{Stamdard\ Error}$ provides an interval estimate

that gives a range of plausible values for the **population proportion**.

Since the probability level used to determine the **critical value** is called the **confidence level**, this interval estimate is formally known as a **confidence interval**.

One more thing: Since the standard error depends on the **population proportion (p)** which is **unknown**, we need to replace **p** by **the sample proportion ($\bar{p}$)**.

$$\sqrt{\frac{p(1-p)}{n}} \rightarrow \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Then the confidence interval is written as:

$$\bar{p} \pm \underbrace{z_c}_{critical\ value} \times \underbrace{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}}_{Stamdard\ Error}$$