CONCEPT: Consider Bernoulli trials: outcomes are 1 or 0 for success or failure, trials are independent, identically distributed, and probability of success for each trial is p (and probability of failure is $q \equiv 1 - p$).

Assume we perform n trials and observe x successes.

Can we construct a confidence interval for the value of p?

If n is large, the answer is a qualified yes, since the central limit theorem guarantees that we can eventually approximate the distribution of x as a normal distribution with known mean and variance.

$$\mu = np \qquad \sigma^2 = npq$$

Note that the mean and variance are exact, although the distribution of x is not.

The rule of thumb is that $n > 30$ is sufficient (though some references say $n > 5$ is enough) unless p is too close to zero or one.

What if n is small? Can we construct a valid confidence interval at a given significance level, α?

The T variable is the key to creating a confidence interval when we have normally distributed X_i:

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

For Bernoulli trials, we proceed in similar fashion. First, we define the sample mean and sample variance as usual:

$$\overline{x} \equiv \frac{x}{n} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}[n\overline{x}(1-\overline{x})^2 + n(1-\overline{x})(0-\overline{x})^2]$$

Some simplification of the sample variance:

$$s^2 = \frac{1}{n-1}[n\overline{x}(1 - 2\overline{x} + \overline{x}^2) - n\overline{x}^3] = \frac{1}{n-1}[n\overline{x}(1-2\overline{x}]$$

The binomial distribution gives the value of $P(X)$.

$$P(X = x) = {}_nC_x p^x q^{n-x} = {}_nC_x p^x (1-p)^{n-x}$$

Defining a Y variable that is analogous to T, we can write a probability distribution.

$$P\left(Y = \frac{\bar{X} - p}{s/\sqrt{n}}\right) = \begin{cases} {}_nC_m p^m (1-p)^{n-m} & y \cdot s/\sqrt{n} + p = \frac{m}{n} \text{ for } m = 0,\ldots,n \\ 0 & \text{otherwise} \end{cases}$$

The problem is that the distribution of Y still depends on p. The secret behind the Z variable is that it has a standard normal distribution. The distribution of Z is the same, regardless of μ and σ.

Even if we $\sigma = \sqrt{npq} = \sqrt{\mu q}$ to define a variable analogous to Z, the problem persists.

$$P\left(W = \frac{\bar{X} - p}{\sqrt{\mu q}/\sqrt{n}}\right) = \begin{cases} {}_nC_m p^m (1-p)^{n-m} & w\sqrt{n\mu q} + \mu = m = 0,\ldots,n \\ 0 & \text{otherwise} \end{cases}$$

Fortunately, we can still do a one-sided hypothesis test on a particular value of p_0. For example:

H_0: $p = p_0$ (the Null hypothesis)
H_A: $p < p_0$ (the Alternative hypothesis)

We use the binomial distribution to calculate $P_0 = P(X \leq x \mid p = p_0)$ where x is the observed number of successes in n trials. For significance level α, we reject H_0 if $P_0 < \alpha$.

We can also do a 2-sided confidence interval, but we have to look at the probability that a low x value has probability less than $\alpha/2$ or that a high x value has probability less than $\alpha/2$.