

Here's a detailed explanation of why the standard error of a percentage (proportion) is: $\sqrt{\frac{p(1-p)}{n}}$...

1. First, we need to recall the definitions of **expected value** (mean) and **variance**:

Expected Value: $E(x) = \sum xP(x)$ = sum of (each possible x-value) x (probability of obtaining that x-value)

Variance: $Var(x) = E(x^2) - [E(x)]^2 = \sum x^2 P(x) - (\sum xP(x))^2$

2. Then we need to realize that proportions follow a binomial distribution.

Binomial Distribution: A) A series of n independent trials
B) Each trial has to have only 2 possible outcomes: 0 or 1
C) Each trial must have a constant probability of being either 0 or 1

Suppose we're interested in the proportion of items that are defective from a manufacturing process.

B) Each item has 2 possible outcomes: (non-defective = 0) or (defective = 1)

We don't know how many items are defective, but we can say that a proportion p of them are defective.

We can randomly select one item from the manufacturing process and see if it is defective or not.

C) Each item has a constant probability of being defective. $P(\text{defective}) = P(x=1) = p$

If we randomly sample n of these items, then each independent item has the same probability of being defective.

A) A series of n independent trials

So if X = (total number of items that are defective from our sample), then X is a binomial random variable.

We are not interested in X (the total number of defective items); we're interested in the *proportion* of defects.

Therefore, we're interested in the variable $\hat{p} = \frac{X}{n} = \frac{\text{the number of defects}}{\text{the number of items we sample}}$

3. Now we can calculate the expected value and variance of our proportion. First, let's calculate the expected value and variance of X (the total number of defects). Remember that the probability of being a defect is p or $P(x=1)$ is p and the probability of not being a defect is $P(x=0) = 1 - p$

Expected Value of one item being defective: $E(x_1) = \sum x_1 P(x_1) = (1)p + (0)(1-p) = p$

The expected number of defects in total is: $E(x) = E(x_1 + x_2 + \dots + x_n) = E(nx_1) = np$

Variance of one item being defective: $Var(x_1) = E(x_1^2) - [E(x_1)]^2 = (1^2)p + (0^2)(1-p) - p^2 = p - p^2 = p(1-p)$

The variance of n items being defective is: $Var(x) = Var(x_1 + x_2 + \dots + x_n) = Var(x_1) + \dots + Var(x_n) = np(1-p)$

If we know the expected value of X is: $E(x) = np$ and we know that our proportion is $\hat{p} = \frac{X}{n}$, then

we can calculate the expected value of our proportion: $E(\hat{p}) = E\left(\frac{X}{n}\right) = E\left(\frac{1}{n}X\right) = \left(\frac{1}{n}\right)E(X) = \frac{1}{n}np = p$

(because $E(ax + b) = aE(x) + b$)

If we know the variance of X is: $Var(x) = np(1 - p)$ and we know that our proportion is $\hat{p} = \frac{X}{n}$, then

we can calculate the variance of our proportion:

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = Var\left(\frac{1}{n}X\right) = \left(\frac{1}{n}\right)^2 Var(X) = \frac{1}{n^2}np(1 - p) = \frac{p(1 - p)}{n}$$

(because $Var(ax + b) = a^2Var(x)$)

To convert this variance to a standard error, we simply need to take the square root: $SE(\hat{p}) = \sqrt{\frac{p(1 - p)}{n}}$

We also know the general formula for a z-statistic: $Z = \frac{(\text{observed}) - (\text{hypothesized})}{\text{Standard Error}} = \frac{(\text{observed}) - (\text{expected})}{\text{Standard Error}}$

So when we run a z-test for proportions, we use: $Z = \frac{p - E(p)}{SE(p)} = \frac{p - \hat{p}}{\sqrt{\frac{p(1 - p)}{n}}}$

That's the formal mathematical reason why it all works out. There is a slightly more intuitive way of understanding this.

When we calculate a proportion, we count the number of items that possess a certain trait (the number of defective items). This means we are running a Bernoulli process – each item has a constant percent chance of being defective.

The variance of a Bernoulli random variable is $Var(x) = p(1-p)$

Whenever we convert a variance to a standard deviation, we simply take the square root. Thus $SD(x) = \sqrt{p(1 - p)}$

The Central Limit Theorem tells us that to convert a standard deviation to a standard error, we divide by the square root of

the sample size: $SE = \frac{SD}{\sqrt{n}}$

Therefore, $SE = \frac{SD}{\sqrt{n}} = \frac{\sqrt{p(1 - p)}}{\sqrt{n}} = \sqrt{\frac{p(1 - p)}{n}}$