

Using Z-scores to test hypotheses about risks.

Suppose one interviewed all 345 of the fraternity and sorority members at Podunk University, and found that 72 of them were smokers. Overall, 19% of Podunk University undergraduates smoke, and 27% of undergraduates in the state smoke. (Assume that the number of Greeks is relatively small compared to the total size of the student body).

- a. Do Podunk U Fraternity and Sorority members smoke more or less than do Podunk University students in general?
- b. Do Podunk U Fraternity and Sorority smoke more or less than do students in the state in general?

### Introduction

Often in risk analysis we want to evaluate whether the frequency of some condition in a population (in this case, the number of undergraduate smokers in the Fraternities and Sororities at a university, but we could also think about the number of leukemia cases in a small town, or the number of automobile accidents in a large city) differs from some expected rate (undergraduate smokers in general, the national leukemia rate, or the automobile accident rate in all large cities in the US or world). The Z-score is a useful tool for this purpose.

- a. A quick calculation show us that  $72/345 = 0.21$ , or 21% of Podunk U Fraternity and Sorority members smoke. This is higher than the university wide rate...or is it? The problem is that the numbers are so close that it's difficult to say whether the observed difference is because of a real underlying difference in the number of smokers or if it's just a question of chance. That is to say, if we asked 345 students at random whether they smoked, how likely is it that we would find 76 smokers? Before we way that the rates differ, we'd like to be able to calculate this value.

What does the Z-score do?

A Z-score compares some observed occurrence of a condition to the expected rate or value of that condition, as normalized to the rate and the sample size.

In plainer English:

A Z-score (see equation 1a and b) has several elements.

The **numerator** (upper number, see equation 1a) compares the value of the observed frequency of the condition to the expected frequency of the condition. The observed frequency is 0.21, and the expected frequency is 0.19. So the numerator is

$$p(\text{observed}) - p(\text{expected}) = 0.21 - 0.19 = 0.02$$

Intuition: When the observed frequency is exactly the same as the expected frequency, the Z-score will be zero, regardless of the value of the denominator. So, a Z-score around zero indicates that the sample is randomly drawn from the larger population.

As the observed frequency gets farther and farther from the expected frequency (in either direction), the absolute value of the Z-score gets bigger.

The **denominator** adjusts the Z-score in two ways: first, it accounts for the number of individuals or cases in the sample, and second, it compares the probability of the expected effect versus the probability of that effect not occurring.

Intuition: as the sample number goes up, we would expect to be more comfortable believing that it provides representative values. An n of two and we really have little idea what the next number will be. As n gets bigger, the absolute value of the Z score gets bigger as the sample number gets bigger.

Less intuitive: as p (which is bounded by 0 and 1) gets farther away from 0.5, the denominator gets smaller, so the Z-score goes up.

The value of the denominator in this problem is

$$\sqrt{\frac{p \times (1-p)}{n}} = \sqrt{\frac{0.19 \times (1-0.19)}{345}} = 0.021$$

So, the Z-score is

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1-p)}{n}}} = Z = \frac{0.02}{0.021} = 0.95$$

Looking at a Z table (such as that on page 393), we find that this is associated with a 0.8289. One way to interpret this is that if we randomly selected 345 students from Podunk University, we would expect to find 65 or 66 smokers (65.55...but we can't actually have

fractions of people!), but there's about a  $100 - 83\% = 17\%$  chance that we would find 72 or 73 smokers (or, for that matter, 57 or 58 smokers).

b. Now we can use the same method to compare these Fraternity and Sorority members with all undergrads in the state:

We already calculated the  $p$ , and it appears that the smoking rate is lower among PU Fraternity and Sorority members than all undergrads in the state. We can test this:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1-p)}{n}}} = \frac{0.21 - 0.27}{\sqrt{\frac{0.27 \times 0.73}{345}}} = -2.5$$

This is associated with 0.0062, so we might say that there is only a 0.62% chance that we would find 72 smokers in a group of 345 undergrads drawn randomly from around the state.

Which is associated with

Equation 1a:  $Z = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1-p)}{n}}}$  for frequencies or

Equation 1b:  $Z = \frac{\hat{x} - np}{\sqrt{n \times p \times (1-p)}}$  for occurrences.

Note that  $\hat{x} = n\hat{p}$

Copyright 2002 David M. Hassenzahl