Inferential Statistics

The objective of research is to discover relations among variables. For example, you might believe that eating spinach makes you stronger (Popeye Hypothesis). One way to examine this hypothesis would be to (1) measure the strength of every person in the world (population), (2) have them eat some spinach, and then (3) measure their strength a second time.

Obviously, we would have difficulty measuring the entire population! So, we select a small group of representative people (sample) and examine the effect of spinach on them by conducting 1 through 3 above. The problem we now face is determining whether what we find in our small sample applies to the larger population.

This is what we are attempting by using inferential statistics.

One difficulty in making this decision is that people vary in strength – some are stronger (or weaker) than the average person. The graph below illustrates a common distribution (normal distribution) of scores for any characteristic, in this case, strength.

The abscissa (horizontal X-axis) reflects relative strength with “0” reflecting the average or mean, positive numbers indicating people who are stronger than average, and negative numbers indicating people who are weaker than average. The height of the curve on the ordinate (vertical Y-axis) reflects the number of people (frequency) who have that level of strength. As you can see, most people tend to cluster near the mean. The further you go from “0” in either direction, the fewer people you find (decreasing frequency).

![Normal Distribution](image)
In fact, in this normal distribution of scores:

1) 68% of the people will be between -1 and +1 (these numbers reflect ± 1 standard deviations from the mean)
2) 95% of the people will be between -2 and +2 (± 2 standard deviations)
3) 99.7% of the people will be between -3 and +3 (± 3 standard deviations)

If we select samples of people, the distribution of any one of these samples will also have a normal distribution. Most of the samples will be very close to the population mean, especially if they are selected randomly. But, a small percentage of these samples may be near the ends of the distribution (e.g., a sample that consists of Olympic weight lifters).

If we wanted to test our Popeye hypothesis… we could select a sample of people, give them spinach, and then measure their strength. Imagine that we find that the mean strength of the people in our sample after eating spinach is greater than the population mean (+ number).

How do we know whether (1) the people in the sample began with average strength and eating spinach made them stronger or (2) the people in the sample just happened to be a little stronger than average and that the spinach really had no effect. The answer is that we cannot know for certain which of these two are correct. Because we cannot know for certain whether (1) or (2) is correct, we use inferential statistics to calculate the likelihood, or probability, that a sample(s) came from a single population.

For instance, if the mean strength of our sample is “+0.5” after eating spinach, we know that there is a good probability that our sample simply contains people from the population that are slightly stronger than normal. In this case, we would tentatively conclude that spinach does not affect strength (but, we must recognize that this conclusion could be wrong because, for example, the natural strength of our sample could have been a little weaker than normal and eating spinach made them a little stronger than normal, a larger shift in strength from the mean than the +0.5 would indicate).

Alternatively, if the mean strength of our sample is “+3.0” after eating spinach, we know that there is only a small probability that our sample came from the “normal” population of non-spinach eaters. In this case, we would tentatively conclude that spinach does make people stronger because our sample now reflects a different population – strong, spinach-eating people (but again, we must recognize that this conclusion could be wrong because, for example, the natural strength of our sample may have been very high to begin with and spinach had no effect).

This is the basic idea underlying inferential statistics. You should understand, however, that most inferential statistics are a little more complicated because you do not know the population parameters (i.e., mean and standard deviation of the population) as described in the previous example. So, in most types of inferential statistics, you are trying to compare two (or more) samples to determine whether they came from the same population or different populations. To put it another way, we are trying to determine if
(1) there is a statistical difference between the means of the samples and, (2) what is the probability that this difference is due to chance? Is there a real effect for the consumption of spinach on resulting strength or is it simply an artifact of our sample?

As discussed above, we make our conclusions based on probabilities, so there is always a chance that we are incorrect. In fact, there are four possible outcomes for any decision. Before discussing these outcomes, you need to understand one final idea. When we conduct inferential statistics, we do not actually test our hypothesis (e.g., spinach increases strength). We actually test the null hypothesis, that there is no effect (e.g., spinach does not affect strength). If we accept the null hypothesis, then we are concluding that spinach does not affect strength. If we reject the null hypothesis, then we are concluding that spinach does affect strength. The four possible outcomes for any decision are illustrated in the following table:

<table>
<thead>
<tr>
<th>What is actually True</th>
<th>Null True (spinach does not change strength)</th>
<th>Null False (spinach actually changes strength)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject Null Hypothesis (assume spinach affects strength)</td>
<td>Type I Error</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>Do not reject Null Hypothesis (assume spinach does not affect strength)</td>
<td>Correct Decision</td>
<td>Type II Error</td>
</tr>
</tbody>
</table>

When making our decision, we can actually control our chance of making a Type 1 error. We do this by setting the probability of obtaining a given result if the null hypothesis is true. This is called the alpha level and in psychology it is typically set at 5%. Recall that any time we select a sample from the population, there is a small chance that we could get a sample of people who are on average much stronger (or weaker) than the population. An alpha level of 5% means that if we ran our spinach experiment 100 times, and spinach does not affect strength, in five of these 100 experiments we would (falsely) conclude that spinach affects strength.

In conclusion, because we cannot include every person in the population in our experiment, we must use a sample and base our decision on probabilities. This means that a certain percentage of the time, we are going to be incorrect. Our goal is to minimize that percentage.