What is Statistics?

These studies include:

-
-
-

To study data, we look at its distribution – what values it takes and how often it takes the values – which we often plot. Some specific distributions are:

-
-
-

We can see skewness from the plot of a distribution. As an example, look at the normal distribution:

When looking at distributions, we are often most interested in examining the following:

-
-
-

One of the most common distributions which occurs in Statistics is the normal distribution (also called the Gaussian distribution or standard normal curve). We often write $N(\mu, \sigma)$ for the normal distribution with mean μ and standard deviation σ.

How do we begin discussing the probability of specific events coming from a distribution? One way is using the standardized value (also called the z-score):

However, we are not often looking at an entire population. Rather we are looking at data from a sample of a given population. What is true for the *entire* population need *not* be true about the sample. If we take a SRS of size n from a population with mean μ and standard deviation σ, the mean and standard deviation of the samples are:

In fact, if the sample size is large the sampling distribution is approximately normal with distribution $\mathbf{N}\left(\mu, \frac{\sigma}{\sqrt{\mathbf{n}}}\right)$. [This is precisely the Central Limit Theorem.] We can even use normal curves (under certain conditions) to approximate samples from a binomial $\mathbf{B}(\mathbf{n}, \mathbf{p})$ distribution: given a SRS of size n from a large population having success p, then

But often we do not know the mean of the population we are examining. However since we know the underlying distribution, we can use information from a SRS to give estimations *with error* for the mean of the underlying population. This is precisely the notion of confidence intervals.

Graphically, we can represent and compare confidence intervals as follows:

To reduce the margin of error, we can:

-
-
-

In fact, this last method can tell us how to choose our sample size:

The idea of confidence intervals allows us to give a method of testing the truth of a hypothesis against observed data. This is the idea of significance testing.

Of course, statistical inference must be implemented carefully. There are many things to consider, especially with the experimental design:

Of course, we could accept/reject H_0 when H_0 is true/false. This results in our statistical inferences being right/wrong. We need a way of measuring the likelihood of this occurring.

		Truth about the population	
		H_0 true	H_a true
Decision based on sample.	Reject H_0		
	Accept H_0		

We need a way of measuring and discussing these errors.

To increase the power, one could:

-
-
-
-

Now in terms of Type I and Type II errors, we have: