Statistics is the art and science of analyzing data:

> There is an art to asking the right questions, then gathering the right information and summarizing it into the right form. And, there is a science to analyzing the data formally. Most often we will be concerned with the science of data analysis.

Computers have changed the strategy of data analysis:

> In pre-computer days, the emphasis was on getting the maximum information with the minimum number of computations. Today, we don't mind doing millions of computations, and constructing detailed graphics, to analyze relatively small data sets. Our goals – analyzing and understanding the data – haven't changed, but our strategy has changed.

Classical statistical methods rely on

1. Strict mathematical assumptions for the underlying populations.
   For example, we may assume that the data are the values of a random sample from a Poisson distribution or from a normal distribution.
2. Asymptotic (large sample) approximations.
   For example, we may use the normal approximation to the sample mean if we feel that the sample size n is large enough.

Modern statistical methods use computing power to

1. Check if classical assumptions are reasonable.
2. Carry out classical analyses, using approximate and exact methods.
3. Carry out valid analyses with no classical counterpart.
4. Use computer graphics for both display and analysis.

Hypothesis Testing Example. In a classic experiment on plant growth, Charles Darwin took 15 pairs of the plant *Zea mays*, where the two plants in each pair were

> *of exactly the same age, were subjected from the first to last to the same conditions, were descended from the same parents.*

One individual was cross-fertilized (CF), the other was self-fertilized (SF). Darwin hypothesized that cross-fertilized plants produced taller offspring than self-fertilized plants.

The heights of the 15 pairs were then measured. The table below lists the heights in inches, along with differences (cross-fertilized minus self-fertilized).

CF:	23.500	12.000	21.0	22.0	19.125	21.500	22.125	20.375
SF:	17.375	20.375	20.0	20.0	18.375	18.625	18.625	15.250
CF−SF:	6.125	−8.375	1.0	2.0	0.750	2.875	3.500	5.125

CF:	18.25	21.625	23.25	21.0	22.125	23.0	12.0
SF:	16.50	18.000	16.25	18.0	12.750	15.5	18.0
CF−SF:	1.75	3.625	7.00	3.0	9.375	7.5	−6.0

The mean difference is 2.61667 inches. In addition, 13 of 15 differences are positive. How significant are these results?

1. *Paired t Test:* If we are willing to assume that the differences data are the values of a random sample from a normal distribution with mean μ, then we could conduct a paired t test of the null hypothesis $\mu = 0$ versus the one-sided alternative $\mu > 0$.

   The observed value of the t statistic is 2.148, and the observed significance level is

   $$p \text{ Value} = P(T \geq 2.148) = 0.02485,$$

   where T is a Student t random variable with 14 df.

2. *Fisher Symmetry Test:* Fisher argued that if the cross-fertilized and self-fertilized seeds are random samples from identical populations, and if their sites (in growing pots) were assigned to members of pairs independently at random, then the 15 differences were equally likely to have been positive or negative. One way to judge significance, he argued, was to compare the sum of differences for the original data to the distribution of sums that we would get by randomly reassigning plus and minus signs to the original differences:

   $$S = \pm 6.125 \pm 8.375 \pm 1.0 \pm 2.0 \pm 0.750 \pm 2.875 \pm 3.500 \pm 5.125$$
   $$\pm 1.75 \pm 3.625 \pm 7.00 \pm 3.0 \pm 9.375 \pm 7.5 \pm 6.0$$

   where either a plus sign or a minus sign is chosen in each summand.

   There are $2^{15} = 32,768$ assignments to consider. The observed sum is 39.25, and the observed significance level is

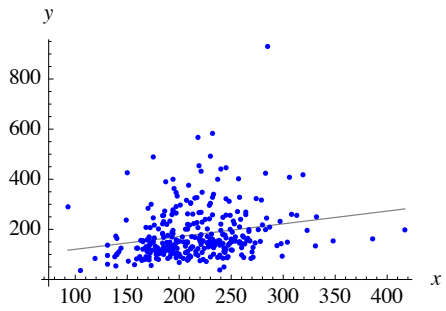
   $$p \text{ Value} = P(S \geq 39.25) = \frac{863}{32768} \approx 0.02634.$$

Footnotes. The paired t test gives trustworthy p values when the data are the values of a random sample from a normal distribution, while Fisher's symmetry test gives trustworthy p values under more general conditions.

Since the p values are close in the example above, both methods yield similar inferences.

Estimation Example. Cholesterol and triglycerides belong to the class of chemicals known as lipids (fats). As part of a study to determine the relationship between high levels of lipids and coronary artery disease, researchers measured plasma levels of cholesterol and triglycerides in milligrams per deciliter (mg/dL) in 371 men complaining of chest pain.

The plot to the right is scatterplot of cholesterol (x) and triglycerides (y) values for the 320 men who showed evidence of disease, superimposed on a least squares fit line. The sample correlation 0.2174.



1. *Fisher's Transformation Method:* If we are willing to assume that the pairs are the values of a random sample from a bivariate normal distribution, then a transformation method developed by Fisher can be used to construct approximate confidence intervals.

   Using Fisher's transformation method, an approximate 95% confidence interval for the correlation between cholesterol and triglycerides is $[0.1103, 0.3194]$.

2. *Efron's Bootstrap Method:* If we are willing to assume that the pairs are the values of a random sample from some bivariate distribution, then a *bootstrap* method developed by Efron can be used. (We pull ourselves up by our own bootstraps.)

   The essence of the method is to use the computer to simulate the original experiment many thousands of times, and to construct approximate confidence intervals based on the results of the simulations.

   Using Efron's method with 5000 simulations, an approximate 95% confidence interval for the correlation between cholesterol and triglycerides is $[0.1173, 0.3105]$.

Footnotes. Finding sample correlations,

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}},$$

and confidence intervals based on the sample correlation, are not usually part of a first course in mathematical statistics. But, we will consider correlation problems this semester.

Confidence interval procedures for pairs drawn from distributions other than the bivariate normal distribution are hard to come by. The bootstrap method is a good alternative approach.

For the example above, the analyses suggest that cholesterol and triglycerides are positively associated in men who experience chest pain and who show evidence of coronary artery disease.