

Statistics at a Glance

Today, we are experiencing an explosion of data collection with sample sizes of ever increasing magnitude and complexity, often called “big data.” Consider the following screen shot of a segment of data collected by the Department of Water Resources in the State of California, USA. It is a graphical representation of the Oroville Dam Reservoir inflow in cfs’s (cubic feet per second) from March 10 through March 12, 2017. The data points were collected at various hourly intervals during that time period.

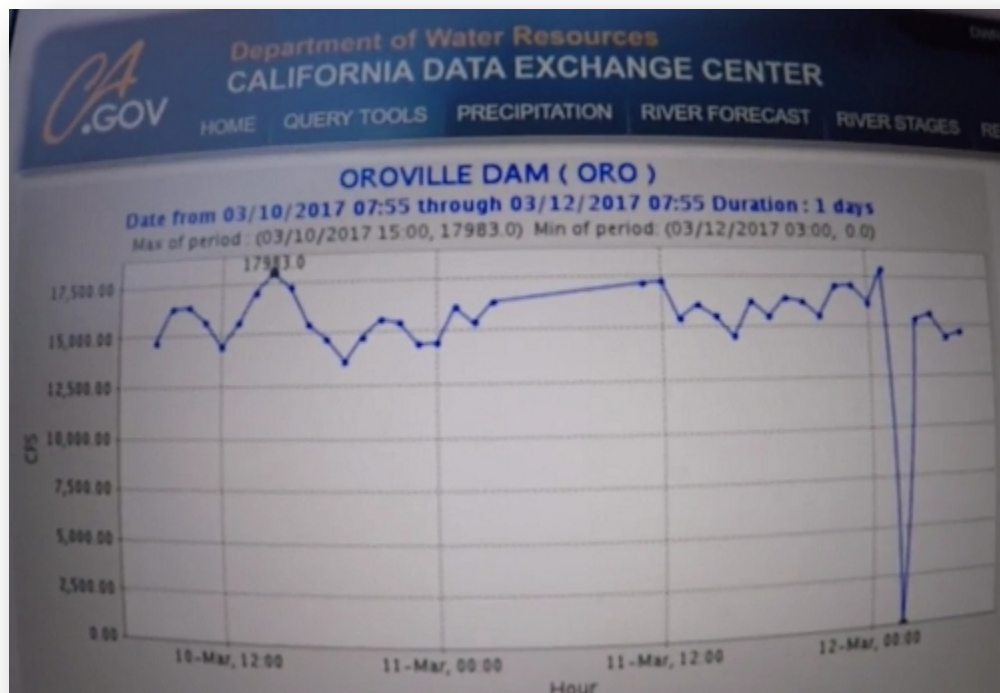


Image Courtesy of: Oroville 12 March update River Valve Outlet System Explained (blancolirio)

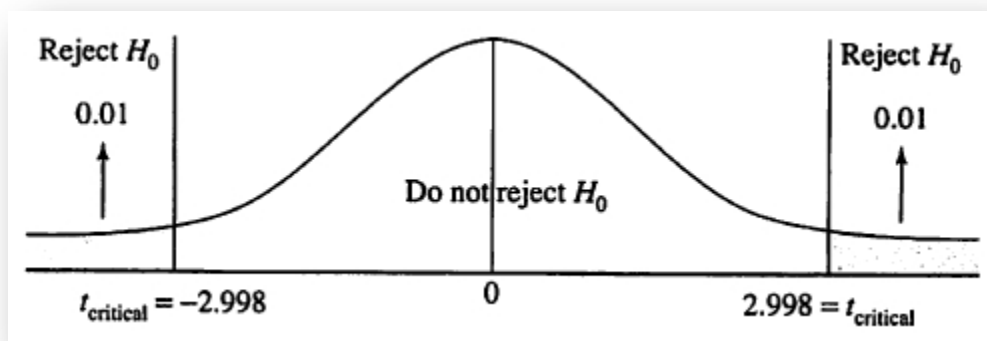
A snapshot of modern-day **descriptive statistics**.

From the data points (sample observations) on this graph, one can determine the maximum, minimum, range (maximum – minimum), mean (average), median (middle point), mode (data value occurring the most), outliers (extreme values) and standard deviation (average difference of the sample measurements from their mean value) of the reservoir inflows over that period of time. All of these terms are quantitative measures which provide some perspective and structure to the data. Those classified as *summary measures* are called “statistics”. Some statistics of particular interest are the mean, the quartiles, the standard deviation and various proportions.

The move from data collected in a relatively small sample of observations about a certain characteristic of some members in a target population to a generalization about that characteristic occurring throughout the entire population is a huge leap of faith. This leap of faith is what we call “**inferential statistics**” and it is supported mathematically by the various theories of the probability of certain kinds of events happening. In hypothesis tests using inferential statistical methods, the results are not guaranteed to be 100% correct all of the time. However, it may be concluded “beyond a reasonable

doubt' (its confidence level) that the results are *probably* valid to its level of confidence. The 'reasonable doubt' is its significance level (often labeled as the first letter of the Greek alphabet, α), which is determined and set by the researcher. A very typical reasonable doubt is 5% but it can go as low as 1% and as high as 10%. This would mean that the corresponding 'beyond a reasonable doubt' values would be 95%, 99% and 90% confident, respectively.

Inferential statistical analysis can be used in a huge variety of situations where the sample observations can be either numerical (quantitative variables) in value or divided into non-numerical, mutually exclusive groups (categorical variables) such as gender, marital status, and voting choices, to name only a few. Below is a snapshot of modern-day **inferential statistics**. Where have the data measurements gone? Statistical processes have abstracted them into a mathematical graph which represents a probability distribution suitable to the kind of data and the nature of the hypotheses.



This is the mathematical abstraction of an actual test.

Because this snapshot is a mathematical abstraction, it could represent a whole array of hypothesis tests with different variable measurements across different disciplines of research BUT which must conform to a certain fixed statistical framework. A lot of information is packed into this image.

From this snapshot alone, one can deduce the following:

The hypothesis test and the number of its tails: It is a two-tailed test because of the two rejection regions. The researcher has hypothesized that the value of the parameter (population characteristic) deduced from the sample observations (H_1) is different (not equal to) the known-to-be-true or generally accepted value of the population parameter (H_0).

The significance level: is the sum of the values (areas) in all rejection regions. In this case, $\alpha = 0.02$ or 2%. This 'reasonable doubt' value is at the discretion of the researcher.

The confidence level: is the value (area) in the non-rejection region. In this case, the confidence level is $1 - \alpha = 0.98$ or 98%. The smaller the significance level, the greater the confidence one may have that the test results (whether H_0 is rejected or not rejected) are valid.

The benchmark probability distribution used in the test: Probability distributions measure the probability (the vertical axis measurement) that a single or a range of *variable values* occur(s) over its entire set of values (the horizontal axis measurement). From the theory of probability, the total of all the

probabilities is 1.0 or 100% because it is absolutely certain that one or other of the variable values has to occur. This implies that the area under the graph of the distribution, being a total of the probabilities of all of its values, is also 1.

In this case, the Student's t distribution has been used. Actually, there are a great many probability distributions used as benchmarks (standard normal, Student's t , Chi-squared, F-distribution, binomial approximation and others), each fitting a particular kind of variable and test.

The type of data variable(s) : Because the Student's t distribution is used, the data variable(s) representing the characteristic(s) of the population under study, though not specified, must be quantitative (numerical), continuous (having decimal representations) and randomly selected from the population members at large.

The implicit assumption of normality or non-normality of the population parameter distribution or whether it matters: Generally this condition must be tested separately. Many statistical hypothesis tests (but not all) require that the probability of a particular characteristic value (population parameter) occurring over the entire population of members has a normal or approximately normal distribution (the shape of a symmetric bell curve) about its population mean value.

However, in this case, because the t distribution was used to test the characteristic under study, if the population was not normally distributed, then the sample used must have been of a sufficient size so that the characteristic observations warranted the conditions of the t distribution benchmark.

The sampling technique: In many hypothesis tests using a benchmark distribution of a continuous variable, samples must be obtained by random sampling – that is, each member (person or object) chosen for the sample measurements must have had an equally likely chance of been selected from the population. There can be no bias in the choices. Otherwise, the results may be skewed or simply false.

The critical values of the test: are determined by the significance level (amount of reasonable doubt), the benchmark distribution used, and the number of tails in the test – *not* from the sample values or their summary statistics. They are the cut-off values which separate the rejection from the non-rejection regions of the distribution for the test. In this case the critical values on the t distribution horizontal axis are 2.998 and – 2.998. The area under the graph between these two values is equal to the confidence coefficient of the test, namely 0.98 out of a total of 1 unit.

The rejection and non-rejection regions: if the *test-statistic* (calculated from the sample itself and the null hypothesized value of the parameter) falls along the horizontal axis of the benchmark distribution within one of the rejection regions, then the null hypothesis is rejected. This implies that the researcher's hypothesis (H_1) probably has some validity beyond the stated reasonable doubt. If the *test-statistic* falls along the horizontal axis within the non-rejection region, then the null hypothesis (H_0) is *not* rejected and this has cast some doubt on the validity of the researcher's hypothesis.

What is *not* evident from the mathematical abstraction of the test are the details of the actual experiment. Every researcher must determine whether his/her project data and the nature of the research question(s) satisfy the above structural framework which will ultimately generate this hypothesized snapshot. Components of that decision-making process include:

1. the **precise details of the experimental design**;
2. the **data variable** of comparison (and whether it is quantitative and continuous);
3. the **population probability distribution** (and whether it is normally distributed);
4. the **sample itself** (and whether it is randomly selected);
5. the **sample size** (and its suitability for the test);
6. the **test statistic** which is determined using *sample* statistics and the null hypothesized value of the parameter;
7. the **p-value of the test** which is determined from the benchmark distribution and the *test statistic*;
8. the **inferred conclusion** which is determined using either the **critical value approach** (by comparing the test statistic to the critical value(s) of the test) or the **p-value approach** (by comparing the p-value of the test to the researcher's significance level).

This inferential process may be summarized as follows:

Data collection and calculations:

Population (person or object) $\Rightarrow$ sample (data collected from some members of the population)

Formulation of the hypotheses

Choice of significance and confidence levels

Data variable(s) + hypotheses $\Rightarrow$ benchmark distribution

Sample data statistics + null hypothesized parameter value $\Rightarrow$ test statistic

Testing using EITHER the critical value approach:

Population characteristic parameter + benchmark distribution + significance level $\Rightarrow$ critical values

Comparison of the test statistic to the critical values $\Rightarrow$ inferential hypothesis conclusion

OR the p-value approach:

Test statistic + benchmark distribution $\Rightarrow$ p-value of the test

Comparison of the p-value to the researcher's level of significance $\Rightarrow$ inferential hypothesis conclusion

At a glance then, a course in introductory statistics consists of two broadly defined areas of study.

1. **Descriptive Statistics:** the various ways to "describe" or summarize succinctly the information in a sample of data, and
2. **Inferential Statistics:** a study of the different methods to test hypotheses. These procedures enable a researcher to test a hypothesis about a particular characteristic obtained from a sample of a targeted population against a generally accepted conclusion about that characteristic of the population itself. In short, these are ways to get from the fill-in-the-blanks researcher's project template to its mathematical abstraction and then on to the inferred conclusion.