Descriptive Statistics

Describe the data using standard methods to determine the average value, the range of data around the average, and other characteristics. The objective of **descriptive statistics** is to communicate the results without attempting to generalize beyond the sample of tests to any other test group. This is an important first step in any analysis. For a reader to understand the basis for the conclusions of any T&E study, an idea of what the data look like is necessary.

Inferential Statistics

The goal of **inferential statistics** is to determine the likelihood that differences in test results could have occurred by chance as a result of the combined effects of unforeseen variables not under the direct control of the tester. It is here the statistical heavy artillery is brought to bear. As a result, most damage to readers of T&E results is inflicted by inferential statistics. You would like to be able to infer the likelihood that the observed results can be generalized to other sets of test results.

Statistical Inference is the process of inferring features of the population from observations of a sample. The analysis addresses the question of the likelihood that an observed difference could have arisen by chance. The *z* test is the simplest example of a statistical test, and it examines the difference between a sample and a population when the variable is a measured quantity.

Forgive a moment's pomposity, but that is really what science is all about: deriving general rules that describe a large class of events, based on observation and experiment of a limited subset of this class. There are at least two issues in this generalization. The first issue is that if the scientist wishes to have confidence in a generalization, he or she must be sure that the people or things chosen to study are representative of the class the scientist wants to describe. This notion leads to the fundamental role of random sampling in the design of experiments. The second issue is that there is always some experimental error associated with the process of measurement; when determining the value of some property of interest, the scientist must also provide a reasonable estimate of the likely error associated with the determination.

No measure of association derived from natural variation in a set of variables, however strong, can establish with certainty that one variable caused another.

Population, Samples, Distribution of Sample Means, and Population Mean

In statistical jargon, you want to make **inferences** about the **population, based on the sample** you have studied. The statistical population has little to do with our everyday notion of population unless we're talking about census data or Gallup polls. The tester's 'population' is the set of desired test results about whom he or she wishes to make generalizations (eg, operational tests of prototypes). In the best of all possible worlds, the tester should sample randomly from this 'population'. In point of fact, this utopia can never be realized, if for no other reason than that any tester does not have the time to test every prototype tens of thousands of times.

The basic idea of all of this is that **the results of any one experiment are influenced by the operation of chance**. If we did the experiment a zillion times and calculated the mean each time, **these mean values would form some distribution centered on the true "population" mean** with an SD equal to the original SD divided by the square root of the sample size. This would seem to imply that the *data* must be normally distributed if we're going to use this approach. Strangely, this is not so.

As it turns out, the **mean values determined from repeated samples of a particular size** are distributed near the true mean in a bell-shaped curve with an **SD equal to the original SD divided by the square root of the sample size**. This new SD, describing the distribution of *mean* values, is called the **standard error** (SE) **of the mean.**

Since the sample means are distributed across a range, and we define a specific range called a **"95% confidence interval"**; that is, we are **95% confident** that the **population mean lies in this interval**. The 95% confidence interval is calculated as follows (the '1.96' comes from the z-table and 2 of these – the plus and minus - capture 95% of the normal distribution area about the sample mean):

95% Confidence Interval: Sample Mean ± 1.96×SE

Putting it another way, because **every measurement is subject to some degree of error**, **every sample mean we calculate will be somewhat different**. Most of the time, the sample means will cluster closely to the population mean, but every so often, we'll end up with a screwball result that differs from the truth just by chance. So if we created a sample based on test results, and compared the results to some desired outcome, we'd have a problem if the one we happened to get was an oddball. What we'd really like to know then is: **If the mean differs from the population mean, is it because of some real difference in the performance, or is it because this is one of those rare times when the tests resulted in some oddball data?** We can never be sure, but statistics tell a lot about how often we can expect the test results to differ *by chance alone.*

There's a key point contained in three little words "of the mean." **Just about everything we do in statistical inference has to do with differences between means.** We use the original data to estimate the mean of the sample and its SE and work forward from there to infer what the entire population is like. It seems that most people who have taken a stats course forget this basic idea when they start to worry about when to use parametric statistics such as *t* tests. Although it is true that parametric statistics hang on the idea of a normal distribution, **all we need is a normal distribution *of the means*, not of the original data**.

Hypothesis Testing

To begin any test, we'll do as all good researchers are supposed to do and state a **null hypothesis**; that is to say, we'll start off by assuming that test results will be no different than the population at large, since we want the new test to continue to yield a system with desired performance. We will then do our darnedest **and hope that we don't reject this hypothesis**. So we phrase the null hypothesis as follows:

Ho: mean of test sample (data group) A = mean of test sample (data group) B

The starting point of the game of statistical inference is almost always to assume that there is no difference between the groups; they are all samples drawn at random from the same statistical population. The next step is to **determine the likelihood that the observed differences could be caused by chance variation alone**. If this **probability is sufficiently small (usually less than 1 in 20), then you "reject" the null hypothesis** and conclude that there is some true difference between the groups. **If so, you are concluding that the test result samples therefore came from different populations, the "alternative hypothesis" (H_1).** That is the meaning behind all those $p < 0.05$'s and $p < 0.0001$'s that appear in the literature. They are simply statements of the probability that the observed difference **could have arisen by chance**.

The **probability or "p" level** associated with any test of significance is only a statement of the likelihood that an observed difference could have arisen by chance. **Of itself, it says nothing about the size or importance of an effect.** Because **probability level is so closely related to sample size**, small effects in large data sets of T&E test efforts can achieve impressive levels of significance. Conversely, T&E efforts involving small numbers of tests may have too little power to detect even fairly large effects.

The **p-value limit for testing statistical significance** is called the **alpha level** — the probability of incorrectly rejecting the null hypothesis—and the resulting error is called, for no apparent reason, a **Type I error**.

There is a **natural link between the confidence interval and hypothesis testing -** if the **confidence interval includes the mean of the null hypothesis, then this is equivalent to not rejecting the null hypothesis**.

Statistical significance is the probability of the observed difference arising by chance. It's really too bad that someone in the history of statistics decided to call this phenomenon "statistical significance" as opposed to, say, "a statistically nonzero effect" or "a statistically present effect" because the term is, somehow, so significant. The basic notion has been perverted to the extent that $p < 0.05$ has become the holy grail of clinical and behavioral research, and that $p < 0.0001$ is cause to close the lab down for the afternoon and declare a holiday. Let's take a closer look at **what determines that magical *p* level**. Three variables enter into the determination of a *z* score (and as we shall see, nearly every other statistical test): (1) the observed difference between means, (2) the SD of the distribution, and (3) the sample size. **A change in any one of these three values can change the calculated statistical significance**. **The bottom line is this: the level of statistical significance—0.05, 0.001, or whatever—indicates the likelihood that the test could have come to a false conclusion**. By itself, it tells you *absolutely nothing* about the actual magnitude of the differences between test groups. Up to now, we haven't talked about practical significance. Basically, this reduces to a judgment (by someone of authority) of how much of a difference might be viewed as practically important enough to warrant some further action.

Sample Size

To be honest, sample size calculations are based on nothing but a hope and a prayer because they require some heroic assumptions about the likely differences you will encounter at the end of the test. In fact, **the only time you are really in a good position to do a defensible sample size calculation is after the test is over** because only then will you have decent estimates of the required parameters. Failing that, it's possible to get about any size of sample you could ever want. As we will show you, every sample size calculation involves separate estimates of four different quantities in order to arrive at the fifth, the sample size. The real talent that statisticians bring to the situation is the ability to fiddle all those numbers through several iterations so that the calculated sample size precisely equals the number of tests you wanted to use in the first place. We take guesses (occasionally educated, more commonly wild) at the difference between means and the SD, fix the alpha and beta levels at some arbitrary figures, and then crank out the sample size.

t-test

The ***t* test is used for measured variables in comparing two means**. The ***unpaired** t test* compares the means of two **independent** samples. The ***paired** t test* compares two paired observations on the **same individual or on matched individuals**.

As a result, many studies involve a comparison of two groups, Concept versus Operational, or prototype A versus prototype B. The statistical analysis of two samples is a bit more complicated than the comparison between a sample and a population. Previously, we used the population SD to estimate the random error we could expect in our calculated sample mean. In a two-sample comparison, this SD is not known and must be estimated from the two samples.

We start off with a **null hypothesis** that the **population values** of the two groups are **not different**. Then we try to show whether or not they are different.

ANOVA

Analysis of variance (ANOVA) allows comparison among more than two sample means. ***One-way ANOVA*** deals with a **single categorical independent variable** (or factor). ***Factorial ANOVA*** deals with **multiple factors** in many different configurations.

The ANOVA actually performs multiple t-tests. However, the use of multiple *t* tests to do two-way comparisons is inappropriate because the process leads to a loss of any interpretable level of significance. What we need is a statistical method that permits us to make a statement about overall differences among tests, following which we could seek out where the differences lie – and this is the ANOVA. Our ANOVA null hypothesis (H_0) and alternative hypothesis (H_1) take the following forms:

H_0: means are equal
H_1: at least one of the means is not equal

So to find the overall effect of the some aspect of a test using ANOVA, we first take the differences **between group means** and the overall mean, square them to get rid of the negative signs, and add them. The sum is then multiplied by the number of objects per group to obtain the **Sum of Squares (between groups)**. The next question is how to get an estimate of the variability **within the groups**. This is done by calculating the sum of the squared differences between individual values and the mean value within each group because this captures individual variability between objects. Because this is based on variation within groups, it is called the **Sum of Squares (within groups)**. The larger the Sum of Squares (between) relative to the Sum of Squares (within), the larger the difference between groups compared to the variation of individual values. However, the Sum of Squares (between groups) contains as many terms as there are groups, and the Sum of Squares (within groups) contains as many terms as there are individual data in all the groups. So the more groups, the larger the Sum of Squares (between), and the more data, the larger the Sum of Squares (within). Since **what we're really trying to do is get the *average* variation between groups and compare it to the *average* variation within groups**, it makes sense to divide the Sum of Squares (between) by the number of groups and divide the Sum of Squares (within) by the number of data. Actually, at this point, a little more sleight of hand emerges. Statisticians start with the number of terms in the sum, then subtract the number of mean values that were calculated along the way. The result is called the **degrees of freedom**, for reasons that reside, believe it or not, in the theory of thermodynamics. Then, dividing the Sum of Squares by the degrees of freedom results in a new quantity called

the **Mean Square**. Finally, the ratio of the two mean squares is a measure of the relative variation between groups of variation within groups and is called an ***F* ratio**. *F* = mean square (between)/mean square (within).

The ***F* test is something like a *t* test; that is, the bigger it is, the smaller the probability that the difference could occur by chance**. And as with a *t* test, you have to look up the value of probability corresponding to the particular value in the back of a book (if the computer hasn't provided it for you).

If all ANOVA had to offer was a small edge over *t* tests in looking after significance levels, it wouldn't be worth all the effort involved in calculating it. But by an extension of the approach, called **factorial ANOVA**, we can **include any number of factors in a single experiment and look at the independent effect of each factor without committing the cardinal sin of distorting the overall probability of a chance difference**. As a bonus, by examining interactions between factors, we can also see whether, for example, some treatments work better on some types of objects or have synergistic effects with other treatments.