Statistics – The Big Picture

Chris Mack, July 2014

The world is full of variation. *Statistics* is a branch of mathematics that seeks to *describe* the variation that is observed, and make *inferences* about a population from measurements made on a sample. Populations are often very large (sometimes infinitely large), and so samples are often the only practical way of gaining insight about the population. A good sample is *representative* of the population, so that there is no bias between descriptions of the sample and of the population. But samples have variability, so they always provide an imperfect representation of the population. In statistics, we seek to quantify the uncertainty that comes from inferences made about a population from a sample.

When deriving a statistical result, we aim for two important and complimentary goals: reliability and generalizability. Reliability is the consistency (repeatability) of the results should the study or experiment be repeated. Generalizability is the ability to apply the results taken from a sample to the general population that the sample represents. To achieve generalizability we need a sample that is independent and free of bias so that it is representative of the population as a whole. To achieve reliability in a statistical analysis, we must make sure that the assumptions that go into the analysis hold true for the data being analyzed.

Types of Data

Statistical analysis begins with data. There are two basic types of data:

Categorical – counts (number of people or occurrences, for example) that fall within each category. The counts are also called the frequency and can also be expressed as a relative frequency (or proportion).

Quantitative – the output of a measurement, usually with units, that represents the amount of something.

Additionally, an *identifier* is a special data type that identifies the subject being measured or counted.

The type of statistical analysis that can be performed depends on the type of data one is analyzing.

Descriptive Statistics

Descriptive statistics provides summary values (each one called a *statistic*) that describe the variation found in data. For categorical data this is usually just the counts or proportions in the various categories, sometimes in the form of a contingency table or a bar chart.

For univariate quantitative data, we plot the variation with a *histogram*, the distribution of data values. We describe this distribution with measures of center (mean or median, for example) and spread (standard deviation or interquartile range, for example) and with descriptions of the shape of the distribution (unimodal vs. bimodal vs. uniform, symmetric vs. skewed, etc.). Sometimes we describe such distributions with models, such as the normal (Gaussian) model.

A *robust statistic* is one where one or a few bad data points (called *outliers*) won't significantly change the value of the statistic. Median and IQR are robust statistics, but mean and standard deviation are not. Thus, whenever a non-robust statistic is being used, we must always be on the look-out for outliers.

Graphs of the data are an essential part of descriptive statistics.

Normal Model

The *normal (Gaussian) model* of a distribution is so important in statistics that it deserves special attention. It is a unimodal and symmetric distribution characterized by two parameters, a mean and a standard deviation. The *standard normal model* has a mean of 0 and a standard deviation of 1, and we call a statistic that has this distribution *z*. We often create these *z* values (also called *z*-scores) by taking the data and subtracting off the mean, then dividing the result by the standard deviation of the sample. Values for the standard normal model are found in tables and readily calculated using spreadsheets and other software. The area under the standard normal curve between two *z*-scores represents the probability of finding data between those *z*-scores.

Inferential Statistics

We often seek to draw conclusions (inferences) about a population from measurements made on a sample. The true value of a population parameter is estimated by a statistic that comes from a sample. The two main types of inferences that we make involve determining *confidence intervals* around the sample statistic and performing *hypothesis testing* on that statistic.

Confidence Intervals – For any statistic that comes from a sample, we can define a confidence interval for that statistic: How confident are we that the true value of the population parameter falls within a certain interval around the sample statistic? This is an expression of our uncertainty in the use of the sample statistic to represent the population parameter, and is equal to our best estimate of the population parameter (the sample statistic) plus or minus the *margin of error*.

Hypothesis Testing – What is the probability (called the *P-value*) that a sample like ours could have come about by *sample variability* given a true null hypothesis? Comparing this P-value to a predefined significance level (α) allows us to reject, or not, the null hypothesis.

The key to both of these types of inferential statistics is knowing the *sampling distribution* for the statistic: If many random samples of a given size were collected and the statistic of interest calculated for each, what would be the resulting distribution of that statistic? The standard deviation of the sampling distribution is a measure of our uncertainty in the statistic of interest. We generally estimate the standard deviation of the sampling distribution using data from our sample, and that estimate is called the *standard error* of the statistic. Generally, this standard error is inversely proportional to the square root of the sample size, so that larger samples produce less uncertainty (a smaller margin of error). Often, the sampling distribution is thought to be normal thanks to the central limit theorem.

The types of hypothesis testing that can be done depend on the type of data, and the sampling distribution we use depends on the specific statistic we are testing.

Hypothesis Testing with Quantitative Data (parametric testing)

One-Sample t-Test: compare the mean of a sample to a hypothesized mean for the population.
Assumptions: The sample data are independent of each other (the sample is random and representative, and less than 10% of the total population), and the sample is large enough (depends on how nearly normal the underlying population is).

Two-Sample t-Test: compare the means that come from two different, independent samples.

Assumptions: The data within each sample are independent of each other (the sample is random and representative, and less than 10% of the total population), the two samples are independent of each other, and each sample is large enough (depends on how nearly normal the underlying population is).

Paired Sample t-Test: compare the mean difference between paired sample data to a hypothesized mean difference for the population. A common example is a before-and-after test.

Assumptions: The data within each sample are independent of each other (the sample is random and representative, and less than 10% of the total population), the two samples are uniquely paired so that the difference has meaning, and the sample is large enough (depends on how nearly normal the underlying population is).

In all of these tests, the sample size must be "large enough" to invoke the central limit theorem so that the sampling distribution of the mean can be assumed to be normal. If the underlying population is very nearly normal, "large enough" is typically 15 – 20. If the underlying population is slightly to moderately skewed, 40 – 50 data points per sample will be required. For heavily skewed or bimodal populations, much larger data sets will be required.

Statistics cannot guarantee certainty. Conclusions from an hypothesis test can be wrong in two ways: a Type I error (rejecting a true null hypothesis) and a Type II error (failing to reject a false null hypothesis).

Hypothesis Testing Procedure

All of the various hypothesis tests that we have used employ the same basic procedure. They all seek to answer this specific question: What is the probability that a sample like this one (or one even more unusual than this one) could have come about by sample variability given a true null hypothesis? If the probability is too low, we reject the null hypothesis. If the probability is high enough, we can't reject the null hypothesis.

Step 0: Plot the data

This step is especially important for two-sample *t*-tests and ANOVA, where boxplots are compared.

Step 1: Check assumptions

The specific assumptions depend on the test being done. All assumptions should be carefully checked, and for quantitative data histograms of the data are required.

Step 2: Write the null and alternate hypotheses

The null hypothesis is generally of the form: H_o: parameter = value

The alternate hypothesis can take one of three forms: H_A: parameter $\neq$ value

H_A: parameter > value

H_A: parameter < value

For the χ^2 and ANOVA tests, a one-tail test is always used (and there is no options with regard to the alternate hypothesis). For proportion and mean tests, the alternate hypothesis looking only for a difference in the parameter value (H_A: parameter $\neq$ value) requires a two-tail test, while the other alternate hypotheses use a one-tail test.

Step 3: Pick a significance level (α)

The default is generally $\alpha = 0.05$, but feel free to change to a different significance level if needed. The significance level is the probability of a Type I error (rejecting a true null hypothesis).

Step 4: Calculate the P-value

The calculation steps here are dependent on the specific test being used, but in general we start by calculating the appropriate test statistic (*z* for a proportion test, *t* for a *t*-test of means, χ^2 for a chi-square test, and *F*-ratio for ANOVA). Then, using the appropriate probability distribution, turn the test statistic into a P-value (the probability of getting data like this, or even more unusual than this, given a true null hypothesis).

Step 5: Form a conclusion about your hypotheses

If P-value < α, reject H_o in favor of the alternate hypothesis. If P-value > α, we fail to reject H_o.

Step 6: If the null hypothesis is rejected, perform additional analysis

Rejecting the null hypothesis means that we have found an effect. We usually wish to put a confidence interval around our estimate of the size of this effect and/or perform an effect size test (such as the Cohen's *d* for *t*-tests and η^2 for ANOVA). For ANOVA, rejecting the null hypothesis means we will also want to perform a Scheffe *post hoc* test to find which mean(s) are different from the others.

Regression Analysis

We often look for relationships between variables. A scatterplot displays the relationship between two quantitative variables, and the Pearson correlation coefficient measures the strength of their *linear* association. A *regression* is a method of finding the 'best' fit of a model to a set of data. A *least-squares linear regression* finds the best-fit line through a set of data under the assumptions that the trend really is linear and that the resulting residuals (data value minus predicted value) are normally distributed with a homogeneous variance. Since least-squares regression is not robust, the results are only valid if there are no outliers. Note that correlations in and of themselves do not imply causation.

The Big Mistakes in Statistics

Unfortunately, there are certain errors in performing statistical analysis or interpreting statistical results that are far too common. Here is a list of the most common errors in the use of statistics.

1. Not properly checking all assumptions inherent to the statistical test being applied.
2. Assuming that all errors are random (and thus ignoring the possibility of systematics errors in measurements or sample bias).
3. Assuming normally distributed data, but not checking the assumption.
4. Using non-robust estimators and statistics without using a reliable procedure for detecting and dealing with outliers.
5. Confusing statistical significance with importance of the effect.
6. Using a test with low power for the desired effect size, so that the null hypothesis is rarely rejected.
7. Trolling for effects: given enough variables, correlations will always be found, whether they exist or not.