Statistics in Brief

Q: What is a p-value?
A: To understand what a p-value is, we need to understand several statistical concepts.

Samples vs Populations.

Statistics are estimates we calculate from a Sample of cases that are representing parameters of all such cases – the True Population. Because the Sample represents the True Population but is not the entire True Population, we feel uncertain that the estimated numbers are true. We express this uncertainty as variability and probability.

Why is our Sample not the True Population? If you study the patients in your clinic, you study them to know how to make them better – and how to make all your future patients better. You want to generalize your results from today for use tomorrow. If you also want to publish your results, meaning you want to teach the world what you learned, then you are also studying your patients to learn how to make all similar patients in the world, current and future, better. The goal of learning something about the "world of similar patients" from a small sample is what makes us need to express our uncertainty about whether our sample truly represents other patients. We measure what we can and infer the Truth from it, giving rise to the term "inferential statistics".

Keep in mind that statistical analysis methods were developed using pencil-and-paper calculations first and then computers later. The ability of computers to execute millions of calculations quickly has permitted statistics to expand greatly. But the basic calculations can be done by hand.

Variability.

Suppose the population mean of some Score is estimated from a large number of observed scores (we can never measure everyone, past present and future). In addition to the mean, we need a measure of variability in that mean, to convey the fact that not everyone scores the same. By how much do they differ? Some scores such as weight are loosely grouped around the mean: weight can vary a lot from person to person. Some scores are tightly grouped around the mean. For example, IQ is tightly grouped around the mean of 100 (potential adult IQ range of about 14-200 or so). If we measure the difference between each person's IQ and the mean of 100, and square those differences (to get rid of negative values), and divide by (n-1) to get an average squared-difference (this number is the Variance), and take the positive square root to get back to the original units of measurement, we have the Standard Deviation: 15 IQ points. Keeping in mind that a particular value may be smaller or greater than the mean, we abbreviate our findings as 100 ± 15 IQ points, or mean 100 (SD=15). The standard deviation describes the average distance between each score and the overall mean. We need to report this measure of variability with the mean to better represent the information.

Here is another person's explanation: http://richardbowles.tripod.com/maths/normdist/normdist.htm

Normal Scores.

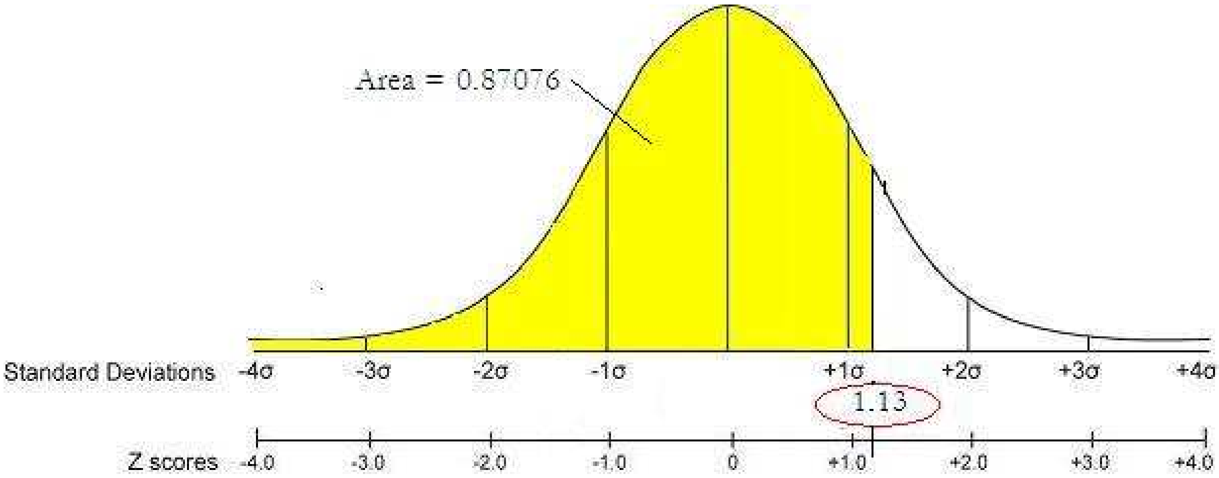
Many of the best-known statistical methods were developed to describe characteristics that are normally distributed continuous measures – measures with lots of numerical values mathematically related to each other, such as weight, IQ score, and blood glucose, measures that have a modal value which is also the median value and the mean value. The mode is the most common or most popular value. The median is the middle value – half your cases are below the median and half your cases are above the median, excluding cases equal to the median. The mean is the average. Imagine a score with negative and positive values, with a mean of 0 and a standard deviation of 1 where the mean, mode, and median are all the same (0). If we graph the scores noting their mean on the x-axis and piling them up like a stack of turtles to show the number of observations on the y-axis, we will get a normal curve. The area under the normal curve will also be 1. The normal curve resembles a diplodocus grazing with its head on the ground, or a boa that has swallowed an elephant as in Le Petit Prince. Here is the pertinent chapter: http://www.angelfire.com/hi/littleprince/chapter1.html And here are graphs of distributions including the normal distribution: http://www.childrensmercy.org/stats/definitions/norm_dist.htm

In statistical analysis, we often want to know whether there is an association between two measures, such as gender and getting treated, gender and weight, or age and weight. A statistical test quantifies the association, and the p-value signals whether we believe the association exists. We compare the pvalue we obtain to some criterion level, usually (alpha=.05), below which we say, "Ah! They are related!" It is important to keep in mind that we may be mistaken. The p-value is part of the evidence that supports or disputes an association. There is always more to the story.

What is the p-value? The p-value is the chance of obtaining the results you get when the null hypothesis is true. The null hypothesis states that there is no association. You are running your analysis to try to find out whether there is an association. If the probability that there is no association is 0.001%, then that is a very slim probability. Such a p-value leads us to think that there is an association. The statistical tests do not actually tell us "there is an association" or "there is no association" because we have uncertainty about our measures and about the Truth. There is always a probability (however slim) that there is no association. In addition, the association we have chosen to test may not be the right one to look at among all the possible associations in the universe. The statistical tests do give us some evidence, however, favoring one hypothesis or another.

The p-value can be calculated several ways but none are easy. For example, p-value can be calculated from normal scores, or z-scores. The z-score compares a measure (an individual observed value) to the norm (the population mean) and then divides that difference by the standard deviation. This is the form of most inferential statistical tests: the difference between the specific and the average, relative to a measure of variability.

If you were comparing a specific score, such as 1.13, from a normally distributed measure to the population mean of 0 (which has a standard deviation of 1), then you could point to the specific score on the normal curve. If you then calculated the area under the normal curve from the left-hand tail up to the specific score, the area under the curve would be the p-value. Here is a picture:



Non-Normal Data.

Suppose you had a measure that was not a normally distributed score. Another set of statistics analyzes dichotomies (presence vs absence; yes vs no), and polytomies (how many were black/white/other?). These types of measures require particular statistical approaches (non-parametric) that differ from the statistics of normally distributed continuous measures. Scales that rank or order values, such as goodbetter- best, are another type of data. Count data (number of days in the hospital; number of outpatient visits) require different approaches as well, because they are not normally distributed. These variations bring up the issue of measurement level.

There are several levels of measurement: ratio, interval, ordinal, and nominal.

Ratio measures have equivalent distance between values and also have a true zero. For example, the temperature measured in degrees Kelvin has a true zero known as absolute zero (the level of energy at which no molecular motion can occur, entropy is minimized, no useful work can be done, there is no randomness). The values between degrees are uniform: we can do math and it will be true that if (7- 5)=5 and (2+3)=5 then (7-5)=(2+3).

Interval measures are similar to ratio measures but have no true zero. Examples include patient weights and blood glucose values. If Patient A changes from 125 lbs to 140 lbs, s/he gains the same number of pounds as Patient B who changes from 200 lbs to 215 lbs. If Patient C changes fasting blood glucose level from 225 microg/dL to 185 microg/dL while Patient D changes from 140 to 100 microg/dL, they have both experienced the same decrease of 40 microg/dL. Note that the mathematical equivalence of these results does not tell us how to interpret the results in terms of patient health. In the blood glucose example, I think we would all agree that the patient reducing his fasting blood glucose to 185 microg/dL has done admirably but is still at risk while the patient whose fasting blood glucose has decreased to 100 does not meet criteria for diabetes. In the case of the two patients gaining 15 pounds, the interpretation depends on how tall and how fit the patients were and why they gained weight (overeating, growing up, bringing a fetus to term?).

Ordinal measures show relationship but do not quantify it. For example, a patient's pain might be "worse today than yesterday" or "better than yesterday" or one of five other descriptive levels of pain. Although we might assign numbers to these levels, we do not know that the intervals are equal. Is the difference between "worse" and "much worse" the same as the difference between "better" and "soso"? Sometimes we can assert that there is an underlying normal distribution and so we can analyze the ordinal scores as if they were interval data. Sometimes it can be shown that this is a false assumption. In the Medical Outcomes Study conducted by RAND, the study that gave us the best-known measures of health-related quality of life – the SF-36 and related scales – the items in the Bodily Pain score were determined to deviate from the expected normal distribution suggested by their numerical values. Therefore, in calculating the Bodily Pain score, these items are weighted before summing.

Finally, nominal or categorical or polytomous measures are unordered. Examples include gender, race, eye color, and political party. We cannot do math with nominal measures: the difference between blue and green cannot be computed or compared in any meaningful way to the difference between brown and grey. Frequently nominal data will be represented by numbers and then subjected to statistical analysis. In these cases, we must always make sure that the analysis does not assume the nominal measures are numerical in a mathematical sense. A special type of nominal data is the dichotomy, such as died (yes vs no), got treatment (True vs False), and usually gender (M vs F).

What Statistical Test Should I Use?

When I wrote that we might want to analyze “gender and getting treated, gender and weight, or age and weight” I mentioned dichotomous measures (gender, getting treated yes vs no) and interval measures (weight, age). When analyzing two or more measures together, we are interested in the associations among the measures.

When there are only two variables, the analysis is unadjusted and bivariate. The most common analysis choices are chi-square analysis for two nominal measures, Student t-test for one dichomous with one interval measure, ANalysis Of VAriance (ANOVA) for one polytomous with one interval measure, and forms of simple regression: linear regression (at least one interval measure for the outcome; any measure for predictor) and logistic regression (at least one dichotomous measure for the outcome; any measure for predictor) **.**

When there are more than two measures, the analysis can be multiple in the outcome (e.g., repeated measures or multiple analysis of variance (MANOVA); time series) or multiple in the predictors (multiple linear or other regression; analysis of co-variance (ANCOVA); Cox proportional hazard survival analysis).

Kaplan-Meier vs Cox.

A Kaplan-Meier curve shows the proportion of cases surviving (y-axis) plotted against the time elapsed. It is equivalent to a Cox survival analysis with no predictors. A Cox proportional hazards model can take covariates (predictors). It is an analytic approach that assesses factors related to a survival curve. Often in published reports of survival analysis, the K-M curve will be depicted followed by a table of covariate effects from a Cox model.

Model Terminology.

Often but not always we think of one measure as the “outcome” and the others as factors or predictors of that outcome. The outcome is the Dependent Variable, also referred to as the left-hand variable or simply “y”, and the factors or predictors are the Independent or right-hand variables or “x”. The left and right come from writing analysis ideas as equations:

LHS = RHS
DV = function of (IV)
Y = f(X)
Treated = f(Gender)
Weight = f(Gender)
Weight = f(Age)
Weight = f(Gender + Age + …)

For a table that relates level of measurement and number of measures to the analytic approach and also includes non-parametric approaches, see: http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm