Appendix A

Basic Statistical Concepts for Sensory Evaluation

Contents

It is important when taking a sample or designing an experiment to remember that no matter how powerful the statistics used, the inferences made from a sample are only as good as the data in that sample. . . . No amount of sophisticated statistical analysis will make good data out of bad data. There are many scientists who try to disguise badly constructed experiments by blinding their readers with a complex statistical analysis.

—O'Mahony (1986, pp. 6, 8)

This chapter provides a quick introduction to statistics used for sensory evaluation data including measures of central tendency and dispersion. The logic of statistical hypothesis testing is introduced. Simple tests on pairs of means (the *t*-tests) are described with worked examples. The meaning of a *p*-value is reviewed.

A.1 Introduction

The main body of this book has been concerned with using good sensory test methods that can generate quality data in well-designed and well-executed studies. Now we turn to summarize the applications of statistics to sensory data analysis. Although statistics are a necessary part of sensory research, the sensory scientist would do well to keep in mind O'Mahony's admonishment: statistical analysis, no matter how clever, cannot be used to save a poor experiment. The techniques of statistical analysis, do however, serve several useful purposes, mainly in the efficient summarization of data and in allowing the sensory scientist to make reasonable conclusions from the information gained in an experiment. One of the most important conclusions is to help rule out the effects of chance variation in producing our results. "Most people, including scientists, are more likely to be convinced by phenomena that cannot readily be explained by a chance hypothesis" (Carver, 1978, p. 387).

Statistics function in three important ways in the analysis and interpretation of sensory data. The first is the simple description of results. Data must be summarized in terms of an estimate of the most likely values to represent the raw numbers. For example, we can describe the data in terms of averages and standard

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5, © Springer Science+Business Media, LLC 2010

deviations (a measure of the spread in the data). This is the descriptive function of statistics. The second goal is to provide evidence that our experimental treatment, such as an ingredient or processing variable, actually had an effect on the sensory properties of the product, and that any differences we observe between treatments were not simply due to chance variation. This is the inferential function of statistics and provides a kind of confidence or support for our conclusions about products and variables we are testing. The third goal is to estimate the degree of association between our experimental variables (called independent variables) and the attributes measured as our data (called dependent variables). This is the measurement function of statistics and can be a valuable addition to the normal sensory testing process that is sometimes overlooked. Statistics such as the correlation coefficient and chi-square can be used to estimate the strength of relationship between our variables, the size of experimental effects, and the equations or models we generate from the data.

These statistical appendices are prepared as a general guide to statistics as they are applied in sensory evaluation. Statistics form an important part of the equipment of the sensory scientist. Since most evaluation procedures are conducted along the lines of scientific inquiry, there is error in measurement and a need to separate those outcomes that may have arisen from chance variation from those results that are due to experimental variables (ingredients, processes, packaging, shelf life). In addition, since the sensory scientist uses human beings as measuring instruments, there is increased variability compared to other analytical procedures such as physical or chemical measurements done with instruments. This makes the conduct of sensory testing especially challenging and makes the use of statistical methods a necessity.

The statistical sections are divided into separate topics so that readers who are familiar with some areas of statistical analysis can skip to sections of special interest. Students who desire further explanation or additional worked examples may wish to refer to O'Mahony (1986), *Sensory Evaluation of Foods, Statistical Methods and Procedures*. The books by Gacula et al. (2009), *Statistical Methods in Food and Consumer Research*, and Piggott (1986), *Statistical Procedures in Food Research*, contain information on more complex designs and advanced topics. This appendix is not meant to supplant courses in statistics, which are recommended for every sensory professional.

It is very prudent for sensory scientists to maintain an open dialogue with statistical consultants or other statistical experts who can provide advice and support for sensory research. This advice should be sought early on and continuously throughout the experimental process, analysis, and interpretation of results. R. A. Fisher is reported to have said, "To call the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to tell you what the experiment died of" (Fisher, Indian Statistical Congress, 1938). To be fully effective, the sensory professional should use statistical consultants early in the experimental design phase and not as magicians to rescue an experiment gone wrong. Keep in mind that the "best" experimental design for a problem may not be workable from a practical point of view. Human testing can necessarily involve fatigue, adaptation and loss of concentration, difficulties in maintaining attention, and loss of motivation at some point. The negotiation between the sensory scientist and the statistician can yield the best practical result.

A.2 Basic Statistical Concepts

Why are statistics so important in sensory evaluation? The primary reason is that there is variation or error in measurement. In sensory evaluation, different participants in a sensory test simply give different data. We need to find the consistent patterns that are not due to chance variation. It is against this background of uncontrolled variation that we wish to tell whether the experimental variable of interest had a reliable effect on the perceptions of our panelists. Unfortunately, the variance in our measurements introduces an element of risk in making decisions. Statistics are never completely foolproof or airtight. Decisions even under the best conditions of experimentation always run the risk of being wrong. However, statistical methods help us to minimize, control, and estimate that risk.

The methods of statistics give us rules to estimate and minimize the risk in decisions when we generalize from a sample (an experiment or test) to the greater population of interest. They are based on consideration of three factors: the actual measured values, the error or variation around the values, and the number

of observations that are made (sometimes referred to as "sample size," not to be confused with the size of a food sample that is served). The interplay of these three factors forms the basis for statistical calculations in all of the major statistical tests used with sensory data, including *t*-tests on means, analysis of variance, and *F*-ratios and comparisons of proportions or frequency counts. In the case of *t*-test on means, the factors are (1) the actual difference between the means, (2) the standard deviation or error inherent in the experimental measurement, and (3) the sample size or number of observations we made.

How can we characterize variability in our data? Variation in the data produces a distribution of values across the available measurement points. These distributions can be represented graphically as histograms. A histogram is a type of graph, a picture of frequency counts of how many times each measurement point is represented in our data set. We often graph these data in a bar graph, the most common kind of histogram. Examples of distributions include sensory thresholds among a population, different ratings by subjects on a sensory panel (as in Fig. A.1), or judgments of product liking on a 9-point scale across a sample of consumers. In doing our experiment, we assume that our measurements are more or less representative of the entire population of people or those who might try our product. The experimental measurements are referred to as a sample and the underlying or parent group as a population. The distribution of our data bears some resemblance to the parent population, but it may differ due to the variability in the experiment and error in our measuring.

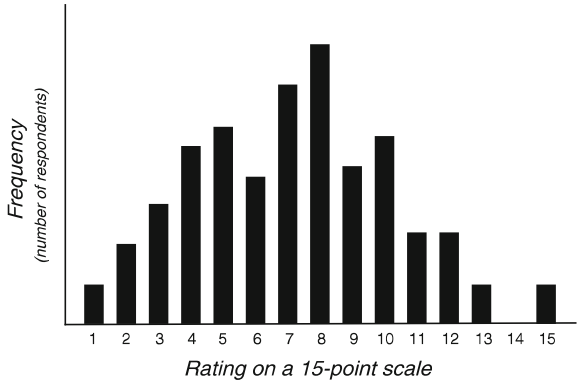


Fig. A.1 A histogram showing a sample distribution of data from a panel's ratings of the perceived intensity of a sensory characteristic on a 15-point category scale.

A.2.1 Data Description

How do we describe our measurements? Consider a sample distribution, as pictured in Fig. A.1. These measurements can be characterized and summarized in a few parameters. There are two important aspects we use for the summary. First, what is the best single estimate of our measurement? Second, what was the variation around this value?

Description of the best or most likely single value involves measures of central tendency. Three are commonly used: the mean is commonly called an average and is the sum of all data values divided by the number of observations. This is a good representation of the central value of data for distributions that are symmetric, i.e., not too heavily weighted in high or low values, but evenly dispersed. Another common measure is the median or 50th percentile, the middle value when the data are ranked. The median is a good representation of the central value even when the data are not symmetrically distributed. When there are some extreme values at the high end, for example, the mean will be unduly influenced by the higher values (they pull the average up). The median is simply the middle value after the measurements are rank ordered from lowest to highest or the average of the two middle values when there is an even number of data points. For some types of categorical data, we need to know the mode. The mode is the most frequent value. This is appropriate when our data are only separated into name-based categories. For example, we could ask for the modal response to the question, when is the product consumed (breakfast, lunch, dinner, or snack)? So a list of items or responses with no particular ordering to the categories can be summarized by the most frequent response.

The second way to describe our data is to look at the variability or spread in our observations. This is usually achieved with a measure called the standard deviation. This specifies the degree to which our measures are dispersed about the central value.

The standard deviation of such an experimental sample of data (S) has the following form:

$$S = \sqrt{\frac{\sum_{i=1}^{N} (X_i - M)^2}{N - 1}} \tag{A.1}$$

where M = mean of X scores = $(\Sigma\ X)/N$.

The standard deviation is more easily calculated as

$$S = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \left((\Sigma X)^2/N\right)}{N-1}} \qquad \text{(A.2)}$$

Since the experiment or sample is only a small representation of a much larger population, there is a tendency to underestimate the true degree of variation that is present. To counteract this potential bias, the value of *N*–1 is used in the denominator, forming what is called an "unbiased estimate" of the standard deviation. In some statistical procedures, we do not use the standard deviation, but its squared value. This is called the sample variance or S^2 in this notation.

Another useful measure of variability in the data is the coefficient of variation. This weights the standard deviation for the size of the mean and can be a good way to compare the variation from different methods, scales, experiments, or situations. In essence the measure becomes dimensionless or a pure measure of the percent of variation in our data. The coefficient of variation (CV) is expressed as a percent in the following formula:

$$\text{CV}(\%) = 100\frac{S}{M} \qquad \text{(A.3)}$$

where *S* is the sample standard deviation and *M* is the mean value. For some scaling methods such as magnitude estimation, variability tends to increase with increasing mean values, so the standard deviation by itself may not say much about the amount of error in the measurement. The error changes with the level of mean. The coefficient of variation, on the other hand, is a relative measure of error that takes into account the intensity value along the scale of measurement.

The example below shows the calculations of the mean, median, mode, standard deviation, and coefficient of variation for data shown in Table A.1.

$N = 41$
Mean of the scores = $(\Sigma X)/N$ = (2 + 3 + 3 + 4 + . . . + 11 + 12 +13) / 41 = 7.049
Median = middle score = 7
Mode = most frequent score = 6
Standard deviation = *S*

Table A.1 First data set, rank ordered

2	5	7	9
3	5	7	9
3	6	7	9
4	6	8	9
4	6	8	10
4	6	8	10
4	6	8	10
5	6	8	11
5	6	8	11
5	7	9	12
			13

$$S = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \left((\Sigma X)^2/N\right)}{N-1}}$$

$$= \sqrt{\frac{2,303 - (83,521)/41}{40}} = 2.578$$

CV (%) = 100 (*S*/mean) = 100 (2.578/ 7.049) = 36.6%.

A.2.2 Population Statistics

In making decisions about our data, we like to infer from our experiment to what might happen in the population as a whole. That is, we would like our results from a subsample of the population to apply equally well when projected to other people or other products. By population, we do not necessarily mean the population of the nation or the world. We use this term to mean the group of people (or sometimes products) from which we drew our experimental panel (or samples) and the group to which we would like to apply our conclusions from the study. The laws of statistics tell us how well we can generalize from our experiment (or sensory test) to the rest of the population of interest. The population means and standard deviations are usually denoted by Greek letters, as opposed to standard letters for sample-based statistics.

Many things we measure about a group of people will be normally distributed. That means the values form a bell-shaped curve described by an equation usually attributed to Gauss. The bell curve is symmetric around a mean value—values are more likely to be close to the mean than far from it. The curve is described by its parameters of its mean and its standard deviation as shown in Fig. A.2. The standard deviation

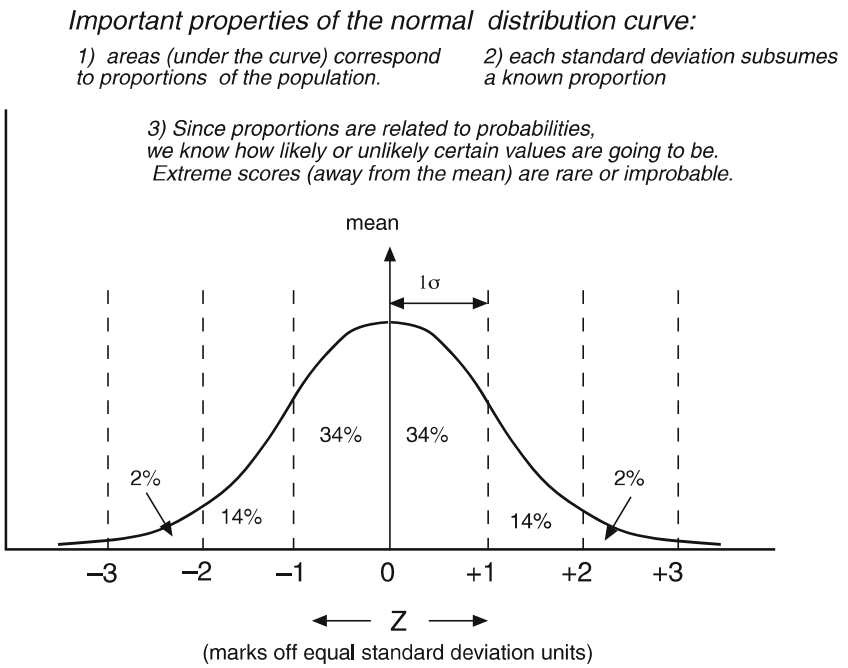


Fig. A.2 The normal distribution curve is described by its parameters of its mean and its standard deviation. Areas under the curve mark off discrete and known percentages of observations.

of a population, σ, is similar to our formula for the sample standard deviation as is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \tag{A.4}$$

where

X = each score (value for each person, product); μ = population mean; N = number of items in population.

How does the standard deviation relate to the normal distribution? This is an important relationship, which forms the basis of statistical risk estimation and inferences from samples to populations. Because we know the exact shape of the normal distribution (given by its equation), standard deviations describe known percentages of observations at certain degrees of difference from the mean. In other words, proportions of observations correspond to areas under the curve. Furthermore, any value, X, can be described in terms of a Z-score, which states how far the value is from the mean in standard deviation units. Thus,

$$Z = \frac{X - \mu}{\sigma} \tag{A.5}$$

Z-scores represent differences from the mean value but they are also related to areas under the normal curve. When we define the standard deviation as one unit, the Z-score is also related to the area under the curve to the left of right of its value, expressed as a percentage of the total area. In this case the z-score becomes a useful value to know when we want to see how likely a certain observation would be and when we make certain assumptions about what the population may be like. We can tell what percent of observations will lie a given distance (Z-score) from the mean. Because the frequency distribution actually tells us how many times we expect different values to occur, we can convert this z-score to a probability value (sometimes called a p-value), representing the area under the curve to the left or right of the Z-value. In statistical testing, where we look for the rarity of calculated event, we are usually examining the "tail" of the distribution or the smaller area that represents the probability of values more extreme than the z-score. This probability value represents the area under the curve outside our given z-score and is the chance (expected frequency) with which we would see a score of that magnitude or one that is even greater. Tables converting z-values to p-values are found in all statistics texts (see Table A).

A.3 Hypothesis Testing and Statistical Inference

A.3.1 The Confidence Interval

Statistical inference has to do with how we draw conclusions about what populations are like based on samples of data from experiments. This is the logic that is used to determine whether our experimental variables had a real effect or whether our results were likely to be due to chance or unexplained random variation. Before we move on to this notion of statistical decision making, a simpler example of inferences about populations, namely confidence intervals, will be illustrated.

One example of inference is in the estimation of where the true population values are likely to occur based on our sample. In other words, we can examine the certainty with which our sample estimates will fall inside a range of values on the scale of measurement. For example, we might want to know the following information: Given the sample mean and standard deviation, within what interval is the true or population value likely to occur? For small samples, we use the *t*-statistic to help us (Student, 1908). The *t*-statistic is like *Z*, but it describes the distribution of small experiments better than the *z*-statistic that governs large populations. Since most experiments are much smaller than populations, and sometimes are a very small sample indeed, the *t*-statistic is useful for much sensory evaluation work. Often we use the 95% confidence interval to describe where the value of the mean is expected to fall 95% of the time, given the information in our sample or experiment.

For a mean value M of N observations, the 95% confidence interval is given by

$$M \pm t\left(S/\sqrt{N}\right) \tag{A.6}$$

where t is the t-value corresponding to N–1 degrees of freedom (explained below), that includes 2.5% of expected variation in the upper tail outside this value and 2.5% in the lower tail (hence a two-tailed value, also explained below). Suppose we obtain a mean value of 5.0 on a 9-point scale, with a standard deviation of 1.0 in our sample, and there are 15 observations. The t-value for this experiment is based on 14 or $n - 1$ degrees of freedom and is shown in Table B to be 2.145. So our best guess is that the true mean lies in the range of $5 \pm 2.145(1/\sqrt{15})$ or between 4.45 and 5.55. This could be useful, for example, if we wanted to insure that our product had a mean score of at least 4.0 on this scale. We would be fairly confident, given the sample values from our experiment that it would in fact exceed this value.

For continuous and normally distributed data, we can similarly estimate a 95% confidence interval on the median (Smith, 1988), given by

$$\text{Med} \pm 1.253t\left(S/\sqrt{N}\right) \tag{A.7}$$

For larger samples, say $N > 50$, we can replace the t-value with its Z approximation, using $Z = 1.96$ in these formulas for the 95% confidence interval. As the number of observations increases, the t-distribution becomes closer to the normal distribution.

A.3.2 Hypothesis Testing

How can we tell if our experimental treatment had an effect? First, we need to calculate means and standard deviations. From these values we do further calculations to come up with values called test statistics. These statistics, like the *Z*-score mentioned above, have known distributions, so we can tell how likely or unlikely the observations will be when chance variation alone is operating. When chance variation alone seems very unlikely (usually one chance in 20 or less), then we reject this notion and conclude that our observations must be due to our actual experimental treatment. This is the logic of statistical hypothesis testing. It is that simple.

Often we need a test to compare means. A useful statistic for small experiments is called Student's *t*-statistic. Student was the pseudonym of the original publisher of this statistic, a man named Gosset who worked for the Guinness Brewery and did not want other breweries to know that Guinness was using statistical methods (O'Mahony, 1986). By small experiments, we mean experiments with numbers of observations per variable in the range of about 50 or less. Conceptually, the *t*-statistic is the difference between the means divided by an estimate of the error or uncertainty around those means, called the standard error of the means.

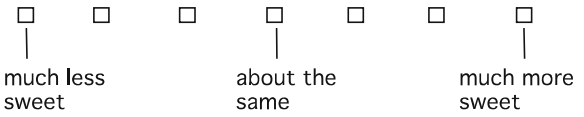
Imagine that we did our experiment many times and each time calculated a mean value. These means themselves, then, could be plotted in a histogram and would have a distribution of values. The standard error of the mean is like the standard deviation of this sampling distribution of means. If you had lots of time and money, you could repeat the experiment over and over and estimate the population values from looking at the distribution of sample mean scores. However, we do not usually do such a series of experiments, so we need a way to estimate this error. Fortunately, the error in our single experiment gives us a hint of how likely it is that our obtained mean is likely to reflect the population mean. That is, we can estimate that the limits of confidence are around the mean value we got. The laws of statistics tell us that the standard error of the mean is simply the sample standard deviation divided by the square root of the number of observations ("*N*"). This makes sense in that the more observations we make, the more likely it is that our obtained mean actually lies close to the true population mean.

In order to test whether the mean we see in our experiment is different from some other value, there are three things we need to know: the mean itself, the sample standard deviation, and the number of observations. An example of this form of the *t*-test is given below, but first we need to take a closer look at the logic of statistical testing.

The logical process of statistical inference is similar for the *t*-tests and all other statistical tests. The only difference is that the *t*-statistic is computed for testing differences between two means, while other statistics are used to test for differences among other values, like proportions, standard deviations, or variances. In the *t*-test, we first assume that there is no difference between population means. Another way to think about this is that it implies that the experimental means were drawn from the same parent population. This is called the null hypothesis. Next, we look at our *t*-value calculated in the experiment and ask how likely this value would be, given our assumption of no difference (i.e., a true null hypothesis). Because we know the shape of the *t*-distribution, just like a *Z*-score, we can tell how far out in the tail our calculated *t*-statistic lies. From the area under the curve out in that tail, we can tell what percent of the time we could expect to see this value. If the *t*-value we calculate is very high and positive or very low and negative, it is unlikely—a rare event given our assumption. If this rarity passes some arbitrary cutoff point, usually one chance in 20 (5%) or less, we conclude that our initial assumption was probably wrong. Then we make a conclusion that the population means are in fact different or that the sample means were drawn from different parent populations. In practical terms, this usually implies that our treatment variable (ingredients, processing, packaging, shelf life) did produce a different sensory effect from some comparison level or from our control product. We conclude that the difference was not likely to happen from chance variation alone. This is the logic of null hypothesis testing. It is designed to keep us from making errors of concluding that the experiment had an effect when there really was only a difference due to chance. Furthermore, it limits our likelihood of making this mistake to a maximum value of one chance in 20 in the long run (when certain conditions are met, see postscript at the end of this chapter).

A.3.3 A Worked Example

Here is a worked example of a simple *t*-test. We do an experiment with the following scale, rating a new ingredient formulation against a control for overall sweetness level:



We convert their box ratings to scores 1 (for the leftmost box) through 7 (for the rightmost). The data from ten panelists are shown in Table A.2.

We now set up our null hypothesis and an alternative hypothesis different from the null. A common notation is to let the symbol H_o stand for the null hypothesis and

Table A.2 Data for *t*-test example

Panelist	Rating
1	5
2	5
3	6
4	4
5	3
6	7
7	5
8	5
9	6
10	4

H_a stand for the alternative. Several different alternatives are possible, so it takes some careful thought as to which one to choose. This is discussed further below. The null hypothesis in this case is stated as an equation concerning the population value, not our sample, as follows:

H_o: $\mu = 4.0$. This is the null hypothesis.
H_a: $\mu \neq 4.0$ This is the alternative hypothesis.

Note that the Greek letter "mu" is used since these are statements about population means, not sample means from our data. Also note that the alternative hypothesis is non- directional, since the population mean could be higher or lower than our expected value of 4.0. So the actual *t*-value after our calculations might be positive or negative. This is called a two-tailed test. If we were only interested in the alternative hypothesis (H_a) with a "greater than" or "less than" prediction, the test would be one tailed (and out critical *t*-value would change) as we would only examine one end of the *t*-distribution when checking for the probability and significance of the result.

For our test against a mean or fixed value, the *t*-test has the following form:

$$t = \frac{M - \mu}{S/\sqrt{N}} \tag{A.8}$$

where M is the sample mean, S is the standard deviation, N is the number of observations (judges or panelists, usually), and μ is the fixed value or population mean.

Here are the calculations from the data set above:

Mean $= \Sigma X/N = 5.0$
$\Sigma X = 50$
$\Sigma X^2 = 262$
$(\Sigma X)^2 = 2500$

$$S = \frac{\sqrt{(262) - (2500)/10}}{9} = 1.155$$

$$t = \frac{5.0 - 4.0}{1.155/\sqrt{10}} = \frac{1}{0.365} = 2.740$$

So our obtained *t*-value for this experiment is 2.740. Next we need to know if this value is larger than what we would expect by chance less than 5% of the time. Statistical tables for the *t*-distribution tell us that for a sample size of 10 people (so degrees of freedom = 9), we expect a *t*-value of ±2.262 only 5% of the time. The two-tailed test looks at both high and low tails and adds them together since the test is non-directional, with *t* high or low. So this critical value of +2.262 cuts off 2.5% of the total area under the *t*-distribution in the upper half and −2.262 cuts off 2.5% in the lower half. Any values higher than 2.262 or lower than −2.262 would be expected less than 5% of the time. In statistical talk, we say that the probability of our obtained result then is less than 0.05, since 2.738 > 2.262. In other words, we obtained a *t*-value from our data that is even more extreme than the cutoff value of 2.262.

So far all of this is some simple math, and then a cross-referencing of the obtained *t*-value to what is predicted from the tabled *t*-values under the null hypothesis. The next step is the inferential leap of statistical decision making. Since the obtained *t*-value was bigger in magnitude than the critical *t*-value, H_o is rejected and the alternative hypothesis is accepted. In other words, our population mean is likely to be different than the middle of our scale value of 4.0. We do not actually know how likely this is, but we know that the experiment would produce the sort of result we see only about 5% of the time when the null is true. So we infer that it is probably false. Looking back at the data, this does not seem too unreasonable since seven out of ten panelists scored higher than the null hypothesis value of 4.0. When we reject the null hypothesis, we claim that there is a statistically significant result. The use of the term "significance" is unfortunate, for in simple everyday English it means "important." In statistical terms significance only implies that a decision has been made and does not tell us whether the result was important or not. The steps in this chain of reasoning, along with some decisions made early in the process about the alpha-level and power of the test, are shown in Fig. A.3.

A.3.4 A Few More Important Concepts

Before going ahead, there are some important concepts in this process of statistical testing that need further explanation. The first is degrees of freedom. When we look up our critical values for a statistic, the values are

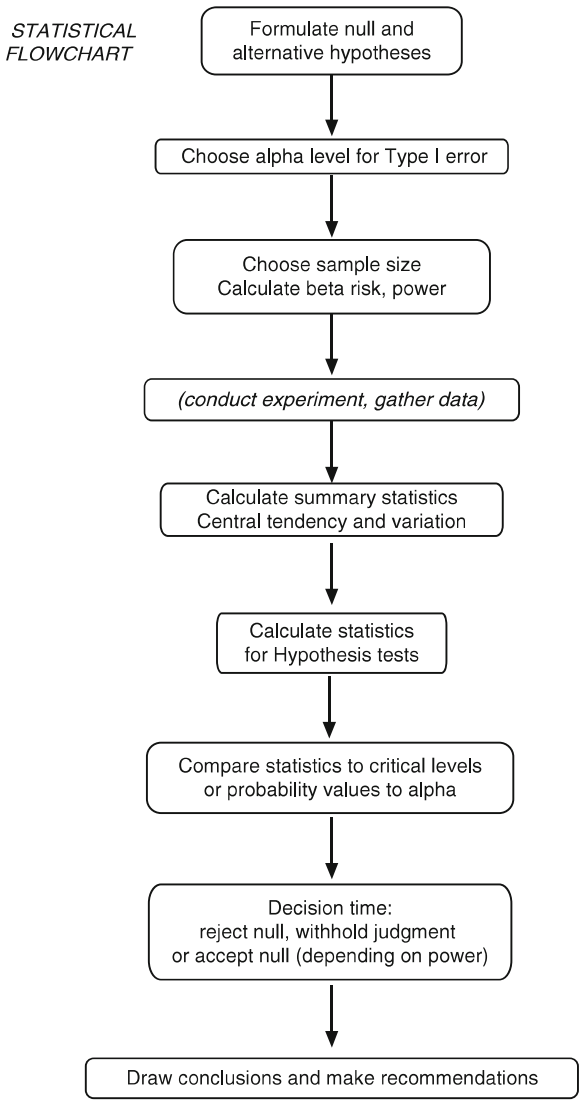


Fig. A.3 Steps in statistical decision making in an experiment. The items before the collection of the data concern the experimental design and statistical conventions to be used in the study. After the data are analyzed the inferential process begins, first with data description, then computation of the test statistic, and then comparison of the test statistic to the critical value for our predetermined alpha-level and the size of the experiment. If the computed test statistic is greater in magnitude than the critical value, we reject the null hypothesis in favor of the alternative hypothesis. If the computed test statistic has a value smaller in magnitude than the critical value, we can make two choices. We can reserve judgment if the sample size is small or we can accept the null hypothesis if we are sure that the power and sensitivity of the test are high. A test of good power is in part determined by having a substantial number of observations and test sensitivity is determined by having good experimental procedures and controls (see Appendix E).

frequently tabled not in terms of how many observations were in our sample, but how many degrees of freedom we have. Degrees of freedom have to do with how many parameters we are estimating from our data relative to the number of observations. In essence, this notion asks how much the resulting values would be free to move, given the constraints we have from estimating other statistics. For example, when we estimate a mean, we have freedom for that value to move or change until the last data point is collected. Another way to think about this is the following: If we knew all but one data point and already knew the mean, we would not need that last data point. It would be determined by all the other data points and the mean itself, so it has no freedom to change. We could calculate what it would have to be. In general, degrees of freedom are equal to the sample size, minus one for each of the parameters we are estimating. Most statistics are tabled by their degrees of freedom. If we wanted to compare the means from two groups of N_1 and N_2 observations, we would have to calculate some parameters like means for each group. So the total numbers of degrees of freedom are N_1-1+N_2-1, or N_1+N_2-2.

A second important consideration is whether our statistical test is a one- or a two-tailed test. Do we wish to test whether the mean is simply different from some value or whether it is larger or smaller than some value? If the question is simply "different from" then we need to examine the probability that our test statistic will fall into either the low or high tail of its distribution. As stated above in the example of the simple *t*-test, if the question is directional, e.g., "greater than" some value, then we examine only one tail. Most statistical tables have entries for one- and two-tailed tests. It is important, however, to think carefully about our underlying theoretical question. The choice of statistical alternative hypotheses is related to the research hypothesis. In some sensory tests, like paired preference, we do not have any way of predicting which way the preference will go, and so the statistical test is two-tailed. This is in contrast to some discrimination tests like the triangle procedure. In these tests we do not expect performance below chance unless there is something very wrong with the experiment. So the alternative hypothesis is that the true proportion correct is greater than chance. The alternative is looking in one direction and is therefore one-tailed.

A third important statistical concept to keep in mind is what type of distribution you are concerned with.

There are three different kinds of distributions we have discussed. First, there are overall population distributions. They tell us what the world would look like if we measured all possible values. This is usually not known, but we can make inferences about it from our experiments. Second, we have sample distributions derived from our actual data. What does our sample look like? The data distribution can be pictured in a graph such as a histogram. Third, there are distributions of test statistics. If the null hypothesis is true, how is the test statistic distributed over many experiments? How will the test statistic be affected by samples of different sizes? What values would be expected, what variance due to chance alone? It is against these expected values that we examine our calculated value and get some idea of its probability.

A.3.5 Decision Errors

Realizing that statistical decisions are based on probabilities, it is clear that some uncertainty is involved. Our test statistic may only happen 5% of the time under a true null hypothesis, but the null might still be true, even though we rejected it. So there is a chance that our decision was a mistake and that we made an error. It is also possible sometimes that we fail to reject the null, when a true difference exists. These two kinds of mistakes are called Type I and Type II errors. A Type I error is committed when we reject the null hypothesis when it is actually true. In terms of a *t*-test comparison of means, the Type I error implies that we concluded that two population means are different when they are in fact the same, i.e., our data were in fact sampled from the same parent population. In other words, our treatment did not have an effect, but we mistakenly concluded that it did. The process of statistical testing is valuable, though, because it protects us from committing this kind of error and going down blind alleys in terms of future research decisions, by limiting the proportion of times we could make these decisions. This upper limit on the risk of Type I error (over the long term) is called alpha-risk.

As shown in Table A.3, another kind of error occurs when we miss a difference that is real. This is called a Type II error and is formally defined as a failure to reject the null hypothesis when the alternative hypothesis is actually true. Failures to detect a difference in a *t*-test or more generally to fail to observe that an experimental treatment had an effect can have important or even devastating business implications. Failing to note that a revised manufacturing process was in fact an improvement would lose the potential benefit if the revision were not adopted as a new standard procedure. Similarly, revised ingredients might be passed over when they in fact produce improvements in the product as perceived by consumers. Alternatively, bad ingredients might be accepted for use if the modified product's flaws are undetected. It is necessary to have a sensitive enough test to protect against this kind of error. The long-term risk or probability of making this kind of mistake is called beta-risk, and one minus the beta-risk is defined as the statistical power of the test. The protection against Type II error by statistical means and by experimental strategy is discussed in Appendix E.

Table A.3 Statistical errors in decision making

		Outcome of sensory evaluation	
		Difference reported	No difference reported
True situation	Products are different	Correct decision	Type II error Prob. is beta-risk
	Products are not different	Type I error Prob. is alpha-risk	Correct decision

A.4 Variations of the *t*-Test

There are three kinds of *t*-tests that are commonly used. One is a test of an experimental mean against a fixed value, like a population mean or a specific point on a scale like the middle of a just-right scale, as in the example above. The second test is when observations are paired, for example, when each panelist evaluates two products and the scores are associated since each pair comes from a single person. This is called the paired *t*-test or dependent *t*-test. The third type of *t*-test is performed when different groups of panelists evaluate the two products. This is called the independent groups *t*-test. The formulas for each test are similar, in that they take the general form of a difference between means divided by the standard error. However,

the actual computations are a bit different. The section below gives examples of the three comparisons of means involving the *t*-statistic.

One type of *t*-test is the test against a population mean or another fixed value, as we saw above in our example and Eq. (A.8). The second kind of *t*-test is the test of paired observations also called the dependent *t*-test. This is a useful and powerful test design in which each panelist evaluates both products, allowing us to eliminate some of the inter-individual variation. To calculate this value of *t*, we first arrange the pairs of observations in two columns and subtract each one from the other member of the pair to create a difference score. The difference scores then become the numbers used in further calculations. The null hypothesis is that the mean of the difference scores is zero. We also need to calculate a standard deviation of these difference scores, and a standard error by dividing this standard deviation by the square root of *N*, the number of panelists

$$t = \frac{M_{\text{diff}}}{S_{\text{diff}}/\sqrt{N}} \tag{A.9}$$

where M_{diff} is the mean of the difference scores and S_{diff} is the standard deviation of the difference scores. Here is an example of a *t*-test where each panelist tasted both products and we can perform a paired *t*-test. Products were rated on a 25-point scale for acceptance. Note that we compute a difference score (*D*) in this situation, as shown in Table A.4.

Table A.4 Data for paired *t*-test example

Panelist	Product A	Product B	Difference	(Difference)^2
1	20	22	2	4
2	18	19	1	1
3	19	17	−2	4
4	22	18	−4	16
5	17	21	4	16
6	20	23	3	9
7	19	19	0	0
8	16	20	4	16
9	21	22	1	1
10	19	20	1	1

Calculations:

sum of $D = 10$, mean of $D = 1$
sum of $D^2 = 68$
standard deviation of $D =$

$$S_{\text{diff}} = \sqrt{\frac{\sum_{i=1}^{N} D_i^2 - \left((\Sigma D)^2/N\right)}{N-1}} = \sqrt{\frac{68 - \left(100/10\right)}{9}} = 2.539\,,$$

and *t* comes from

$$t = \frac{M_{\text{diff}}}{S_{\text{diff}}/\sqrt{N}} = \frac{1.0}{2.5390/\sqrt{10}} = 1.25$$

This value does not exceed the tabled value for the 5%, two-tailed limit on *t* (at 9 df), and so we conclude there is insufficient evidence for a difference. In other words, we do not reject the null hypothesis. The two samples were rather close, compared to the level of error among panelists.

The third type of *t*-test is conduced when there are different groups of people, often called an independent groups *t*-test. Sometimes the experimental constraints might dictate situations where we have two groups that taste only one product each. Then a different formula for the *t*-test applies. Now the data are no longer paired or related in any way and a different calculation is needed to estimate the standard error, since two groups were involved and they have to be combined somehow to get a common estimate of the standard deviations. We also have some different degrees of freedom, now given by the sum of the two group sizes minus 2 or $(N_{\text{Group1}} + N_{\text{Group2}} - 2)$. The *t*-value is determined by

$$t = \frac{M_1 - M_2}{SE_{\text{pooled}}} \tag{A.10}$$

where M_1 and M_2 are the means of the two groups and SE_{pooled} is the pooled standard error. For the independent *t*-test, the pooled error requires some work and gives an estimate of the error combining the error levels of the two groups. The pooled standard error for two groups, *X* and *Y*, is given by the following formula:

$$SE_{\text{pooled}} = \sqrt{\left[1/N_1 + 1/N_2\right] \frac{\left[\Sigma x^2 - \left((\Sigma x)^2/N_1\right) + \Sigma y^2 - \left((\Sigma y)^2/N_2\right)\right]}{(N_1 + N_2 - 2)}} \tag{A.11}$$

Here is a worked example of an independent group's *t*-test. In this case, we have two panels, one from a manufacturing site and one from a research site, both evaluating the perceived pepper heat from an ingredient submitted for use in a highly spiced product.

The product managers have become concerned that the plant QC panel may not be very sensitive to pepper heat due to their dietary consumption or other factors, and that the use of ingredients is getting out of line with what research and development personnel feel is an appropriate level of pepper. So the sample is evaluated by both groups and an independent group's *t*-test is performed. Our null hypothesis is that there is no difference in the population means and our alternative hypothesis that the QC plant will have lower mean ratings in the long run (one-tailed situation). The data set is comprised of pepper heat ratings on a 15-point category scale as shown in Table A.5.

Table A.5 Data for independent group's *t*-test

Manufacturing QC panel (*X*)	R&D test panel (*Y*)
7	9
12	10
6	8
5	7
8	7
6	9
7	8
4	12
5	9
3	

First, some preliminary calculations:

$N_1 = 10 \quad \Sigma x = 63 \quad \text{Mean} = 6.30 \quad \Sigma x^2 = 453 \quad (\Sigma x)^2 = 3969$
$N_2 = 9 \quad \Sigma y = 79 \quad \text{Mean} = 8.78 \quad \Sigma y^2 = 713 \quad (\Sigma y)^2 = 6291$

Now we have all the information we need to calculate the value of

$$SE_{\text{pooled}} = \sqrt{(1/10 + 1/9)\frac{\left[453 - \frac{3969}{10} + 713 - \frac{6241}{9}\right]}{(10 + 9 - 2)}} = 0.97$$

$t = [(6.30–8.78)]/0.97 = -2.556.$

Degrees of freedom are 17 (= 10 + 9 − 2). The critical *t*-value for a one-tailed test at 17 df is 1.740, so this is a statistically significant result. Our QC panel does seem to be giving lower scores for pepper heat than the R&D panel.

Note that the variability is also a little higher in the QC panel. Our test formula assumes that the variance is about equal. For highly unequal variability (1 SD more than three times that of the other) some adjustments must be made. The problem of unequal variance becomes more serious when the two groups are also very different in size. The *t*-distribution becomes a poor estimate of what to expect under a true null, so the alpha-level is no longer adequately protected. One approach is to adjust the degrees of freedom and formulas for this are given in advanced statistics books (e.g., Snedecor and Cochran, 1989). The non-pooled estimates of the *t*-value are provided by some statistics packages and it is usually prudent to examine these adjusted *t*-values if unequal group size and unequal variances happen to be the situation with your data.

A.4.1 The Sensitivity of the Dependent t-Test for Sensory Data

In sensory testing, it is often valuable to have each panelist try all of the products in our test. For simple paired tests of two products, this enables the use of the dependent *t*-test. This is especially valuable when the question is simply whether a modified process or ingredient has changed the sensory attributes of a product. The dependent *t*-test is preferable to the separate-groups approach, where different people try each product. The reason is apparent from the calculations. In the dependent *t*-test, the statistic is calculated on a difference score. This means that the differences among panelists in overall sensory sensitivity or even in their idiosyncratic scale usage are removed from the situation. It is common to observe that some panelists have a "favorite" part of the scale and may restrict their responses to one section of the allowable responses. However, with the dependent *t*-test, as long as panelists rank order the products in the same way, there will be a statistically significant result. This is one way to partition the variation due to subject differences from the variation due to other sources of error. In general, partitioning of error adds power to statistical tests, as shown in the section on repeated measures (or complete block) ANOVA (see Appendix C). Of course, there are some potential problems in having people evaluate both products, like sequential order effects and possible fatigue and carry-over effects. However, the advantage gained in the sensitivity of the test usually far outweighs the liabilities of repeated testing.