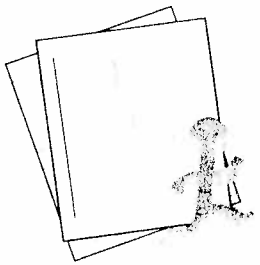
From the book
"JMP Start Statistics",
3rd Ed., by Sall,
Creighton, and Lehmann,
2005, Thomson/Brooks/Cole



What Are Statistics?

Overview

Statistics are numbers, but the practice of statistics is the craft of measuring imperfect knowledge. That's one definition, and there are many more.

This chapter is a collection of short essays to get you started on the many ways of statistical thinking and to get you used to the terminology of the field.

Ponderings

The Business of Statistics

The discipline of statistics provides the framework of balance sheets and income statements for scientific knowledge. Statistics is an accounting discipline, but instead of accounting for money, it is accounting for scientific credibility. It is the craft of weighing and balancing observational evidence. Scientific conclusions are based on experimental data in the presence of uncertainty, and statistics is the mechanism to judge the merit of those conclusions—the statistical tests are like credibility audits. Of course, you can juggle the books and make poor science sometimes look better than it is. However, there are important phenomena that you just can't uncover without statistics.

A special joy in statistics is when it can be used as a discovery tool to find out new phenomena. There are many views of your data—the more perspectives you have on your data, the more likely you are to find out something new. Statistics as a discovery tool is the auditing process that unveils phenomena that are not anticipated by a scientific model and are unseen with a straightforward analysis. These anomalies lead to better scientific models.

Statistics fits models, weights evidence, helps identify patterns in data and then helps find data points that don't fit the patterns. Statistics is induction from experience; it is there to keep score on the evidence that supports scientific models.

Statistics is the science of uncertainty, credibility accounting, measurement science, truth-craft, the stain you apply to your data to reveal the hidden structure, the sleuthing tool of a scientific detective.

Statistics is a necessary bureaucracy of science.

The Yin and Yang of Statistics

There are two sides to statistics.

First, there is the Yang of statistics, a shining sun. The Yang is always illuminating, condensing, evaporating uncertainty, and leaving behind the precipitate of knowledge. It pushes phenomena into forms. The Yang is out to prove things in the presence of uncertainty and ultimately compel the data to confess its form, conquering ignorance headlong. The Yang demolishes hypotheses by ridiculing their improbability. The Yang mechanically cranks through the ore of knowledge and distills it to the answer.

On the other side, we find the contrapositive Yin, the moon, reflecting the light. The Yin catches the shadow of the world, feeling the shape of truth under the umbra. The Yin is forever looking and listening for clues from the creator, nurturing seeds of pattern and anomaly into maturing discoveries. The Yin whispers its secrets to our left hemisphere. It unlocks doors for us in the middle of the night, planting dream seeds, making connections. The Yin draws out the puzzle pieces to tantalize our curiosity. It teases our awareness and tickles our sense of mystery until the climax of revelation—Eureka!

The Yin and Yang are forever interacting, catalyzed by Random, the agent of uncertainty. As we see the world reflected in the pool of experience, the waters are stirred, and in the agitated surface we can't see exactly how things are. Emerging from this, we find that the world is knowable only by degree, that we have knowledge in measure, not in absolute.

The Faces of Statistics

Everyone has a different view of statistics.

Match the definition on this side....	**with someone likely to have said it on this side**
1. The literature of numerical facts.	a. Engineer
2. An applied branch of mathematics.	b. Original meaning
3. The science of evidence in the face of uncertainty.	c. Social scientist
4. A digestive process that condenses a mass of raw data into a few high-nutrient pellets of knowledge.	d. Philosopher
5. A cooking discipline with data as the ingredients and methods as the recipes.	e. Economist
6. The calculus of empiricism.	f. Computer scientist
7. The lubricant for models of the world.	g. Mathematician
8. A calibration aid.	h. Physicist
9. Adjustment for imperfect measurement.	i. Baseball fan
10. An application of information theory.	j. Lawyer
11. Involves a measurable space, a sigma algebra, and Lebesgue integration.	k. Joe College
12. The nation's state.	l. Politician
13. The proof of the pudding.	m. Businessman
14. The craft of separating signal from noise.	n. Statistician
15. A way to predict the future.	

An interesting way to think of statistics is as a toy for grown-ups; remember that toys are proxies that children use to model the world. Children use toys to learn behaviors and develop explanations and strategies, as aids for internalizing the external. This is the case with statistical models. You model the world with a mathematical equation, and then see how the model stacks up to the observed world.

Statistics lives in the interface of the real world data and mathematical models, between induction and deduction, empiricism and idealism, thought and experience. It seeks to balance real data and a mathematical model. The model addresses the data and stretches to fit. The model changes and the change of fit is measured. When the model doesn't fit, the data suspends from the model and leaves clues. You see patterns in the data that don't fit—this leads to a better model, and points that don't fit into patterns can lead to important discoveries.

Don't Panic

Some university students have a panic reaction to the subject of statistics. Yet most science, engineering, business, and social science majors usually have to take at least one statistics course. What are some of the sources of our phobias about statistics?

Abstract Mathematics

Though statistics can become very mathematical to those so inclined, applied statistics can be used effectively with only a very basic level of mathematics. You can talk about statistical properties and procedures without plunging into abstract mathematical depths. In this book, we are interested in looking at applied statistics.

Lingo

Statisticians often don't bother to translate terms like 'heteroschedasticity' into 'varying variances' or 'multicollinearity' into 'closely related variables.' Or, for that matter, further translate 'varying variances' into 'difference in the spread of values between samples,' and 'loosely related variables' into 'variables that give similar information.' We will tame some of the common statistical terms in the discussions that follow.

Awkward Phrasing

There is a lot of subtlety in statistical statements that can sound awkward, but the phrasing is very precise and means exactly what it says. Sometimes statistical statements include multiple negatives. For example, "The statistical test failed to reject the null hypothesis of no effect at the specified alpha level." That is a quadruple negative statement—count the negatives: 'fail,' 'reject,' 'null,' and 'no effect.' You can reduce the negatives by saying "the statistical results are not significant" as long as you are careful not to confuse that with the statement "there is no effect." Failing to prove something does not prove the opposite!

A Bad Reputation

The tendency to assume the proof of an effect because you cannot statistically prove the absence of the effect is the origin of the saying, "Statistics can prove anything." This is what happens when you twist a term like 'nonsignificant' into 'no effect.' This idea is common in a courtroom; you can't twist the phrase "there is not enough evidence to prove beyond reasonable doubt that the accused committed the crime" with "the accused is innocent." What nonsignificant really means is that there is not enough data to show a significant effect—it does not mean that you are certain there is no effect at all.

Uncertainty

Although we are comfortable with uncertainty in ordinary daily life, we are not used to embracing it in our knowledge of the world. We think of knowledge in terms of hard facts and solid logic, though much of our most valuable real knowledge is less than solid. We can say when we know something for sure (yesterday it rained), and we can say when we don't know (don't know whether it will rain tomorrow). But when we describe knowing something with incomplete certainty, it sounds apologetic or uncommitted. For example, it sounds like a form of equivocation to say that there is a 0.9 chance that it will rain tomorrow. Yet much of what we think we know is really just that kind of uncertainty.

Preparations

A few fundamental concepts will prepare you for absorbing details in upcoming chapters.

Three Levels of Uncertainty

Statistics is about uncertainty, but there are several different levels of uncertainty that you have to keep in separate accounts:

Random Events

Even if you know everything possible about the world, you still have events you can't predict. You can see an obvious example of this in any gambling casino. You can be an expert at playing blackjack, but the randomness of the card deck renders the outcome of any game indeterminate. We make models with random error terms to account for uncertainty due to randomness. Some of the error term may be due to ignoring details; some may be measurement error; but much of it is attributed to inherent randomness.

Unknown Parameters

Not only are you uncertain how an event is going to turn out, you often don't even know what the numbers (parameters) are in the model that generates the events. You have to estimate the parameters and test if hypothesized values of them are plausible, given the data. This is the chief responsibility of the field of statistics.

Unknown Models

Sometimes you not only don't know how an event is going to turn out, and you don't know what the numbers are in the model, but you don't even know if the form of the model is right.

Statistics is very limited in its help for certifying that the model is correct. Most statistical conclusions assume that the model is correct. The correctness of the model is the responsibility of the subject-matter science. Statistics might give you clues if the model is not carrying the data very well. Statistical analyses can give diagnostic plots to help you see patterns that could lead to new insights, to better models.

Probability and Randomness

In the old days, statistics texts all began with chapters on probability. Today, many popular statistics books discuss probability in later chapters. We skip the balls-in-urns entirely, though probability is the essence of our subject.

Randomness makes the world interesting and probability is needed as the measuring stick. Probability is the aspect of uncertainty that allows the information content of events to be measured. If the world were deterministic, then the information value of an event would be zero because it would already be known to occur; the probability of the event occurring would be 1. The sun rising tomorrow is a nearly deterministic event and doesn't make the front page of the newspaper when it happens. The event that happens but has been attributed to having probability near zero would be big news. For example, the event of extraterrestrial intelligent life forms landing on earth would make the headlines.

Statistical language uses the term probability on several levels:

- When we make observations or collect measurements, our responses are said to have a *probability distribution*. For whatever reason, we assume that something in the world adds randomness to our observed responses, which makes for all the fun in analyzing data that has uncertainty in it.
- We calculate statistics using probability distributions, seeking the safe position of maximum likelihood, which is the position of least improbability.
- The significance of an event is reported in terms of probability. We demolish statistical null hypotheses by making their consequences look incredibly improbable.

Assumptions

Statisticians are naturally conservative professionals. Like the boilerplate of official financial audits, statisticians' opinions are full of provisos such as "assuming that the model is correct, and assuming that the error is Normally distributed, and assuming that the observations are independent and identically distributed, and assuming that there is no measurement error, and assuming...." Even then the conclusions are hypothetical, with phrases like "if you say the hypothesis is false, then the probability of being wrong is less than 0.05."

Statisticians are just being precise, though they sound like they are combining the skills of equivocation like a politician, techno-babble like a technocrat, and trick-prediction like the Oracle at Delphi.

Ceteris Paribus

A crucial assumption is the *ceteris paribus* clause, which is Latin for other things being equal. This means we assume that the response we observed was really only affected by the model and random error; all other factors that might affect the response were maintained at the same controlled value across all observations or experimental units. This is, of course, often not the case, especially in observational studies, and the researcher must try to make whatever adjustments, appeals, or apologies to atone for this. When statistical evidence is admitted in court cases, there are endless ways to challenge it, based on all the assumptions that may have been violated.

Is the Model Correct?

The most important assumption is that your model is right. There are no easy tests for this assumption. Statistics almost always measure one model against a submodel, and these have no validity if neither model is appropriate in the first place.

Is the Sample Valid?

The other supremely important issue is that the data relate to your model; that is, that you have collected your data in a way that is fair to the questions that you will be asking it. If your sample is ill-chosen, or if you have skewed your data by rejecting data in a process that relates to its applicability to the questions, then your judgments will be flawed. If you have not taken careful consideration of the direction of causation, you may be in trouble. If taking a response affects the value of another response, then they are not independent of each other, which can affect the study conclusions.

In brief, are your samples fairly taken and are your experimental units independent?

Data Mining?

One issue that most researchers are guilty of to a certain extent is stringing together a whole series of conclusions and assuming that the joint conclusion has the same confidence as the individual ones. An example of this is data mining, in which hundreds of models are tried until one is found with the hoped-for results. Just think about the fact that if you collect purely random data, you will find a given test significant at the 0.05 level about 5% of the

time. So you could just repeat the experiment until you get what you want, discarding the rest. That's obviously bad science, but something similar often happens in published studies. This multiple-testing problem remains largely unaddressed by statistical literature and software except for some special cases such as means comparisons, a few general methods that may be inefficient (Bonferroni's adjustment), and expensive, brute-force approaches (resampling methods).

Another problem with this issue is that the same kind of bias is present across unrelated researchers because nonsignificant results are often not published. Suppose that 20 unrelated researchers do the same experiment, and by random chance one researcher got a 0.05- level significant result. That's the result that gets published.

In light of all the assumptions and pitfalls, it is appropriate that statisticians are cautious in the way they phrase results. Our trust in our results has limits.

Statistical Terms

Statisticians are often unaware that they use certain words in a completely different way than other professionals. In the following list, some definitions will be the same as you are used to, and some will be the opposite:

Model

The statistical model is the mathematical equation that predicts the response variable as a function of other variables, together with some distributional statements about the random terms that allow it to not fit exactly. Sometimes this model is taken very casually in order to look at trends and tease out phenomena, and sometimes the model is taken very seriously.

Parameters

To a statistician, parameters are the unknown coefficients in a model, to be estimated and to test hypotheses about. They are the indices to distributions; the mean and standard deviation are the location and scale parameters in the Normal distribution family.

Unfortunately, engineers use the same word (parameters) to describe the factors themselves.

Statisticians usually name their parameters after Greek letters, like mu(μ), sigma(σ), beta(β), and theta(θ). You can tell where statisticians went to school by which Greek and Roman letters they use in various situations. For example, in multivariate models, the L-Beta-M fraternity is distinguished from C-Eta-M.

Hypotheses

In science, the hypothesis is the bright idea that you want to confirm. In statistics, this is turned upside down because it uses logic analogous to a proof-by-contradiction. The so-called *null hypothesis* is usually the statement that you want to demolish. The usual null hypothesis is that some factor has no effect on the response. You are of course trying to support the opposite, which is called the *alternative hypothesis*. You support the alternative hypothesis by statistically rejecting the null hypothesis.

Two-Sided versus One-Sided, Two-Tailed versus One-Tailed

Most often, the null hypothesis can be stated as some parameter in a model being zero. The alternative is that it is not zero, which is called a two-sided alternative. In some cases, you may be willing to state the hypothesis with a one-sided alternative, for example that the parameter is greater than zero. The one-sided test has greater power at the cost of less generality. These terms have only this narrow technical meaning, and it has nothing to do with common English phrases like presenting a one-sided argument (prejudiced, biased in the everyday sense) or being two-faced (hypocrisy or equivocation). You can also substitute the word "tailed" for "sided." The idea is to get a big statistic that is way out in the tails of the distribution where it is highly improbable. You measure how improbable by calculating the area of one of the tails, or the other, or both.

Statistical Significance

Statistical significance is a precise statistical term that has no relation to whether an effect is of practical significance in the real world. Statistical significance usually means that the data gives you the evidence to believe that some parameter is not the null value. If you have a ton of data, you can get a statistically significant test when the values of the estimates are practically zero. If you have very little data, you may get an estimate of an effect that would indicate enormous practical significance, but it is supported by so little data that it is not statistically significant. A nonsignificant result is one that might be the result of random variation rather than a real effect.

Significance Level, *p*-value, α-level

To reject a null hypothesis, you want small *p*-values. The *p-value* is the probability of being wrong if you declare an effect to be non-null; that is, the probability of rejecting a 'true' null hypothesis. The *p*-value is sometimes labeled the *significance probability*, or sometimes labeled more precisely in terms of the distribution that is doing the measuring. The *p*-value labeled "Prob>|t|" is read as "the probability of getting a greater t (in absolute value)." The α-level is your standard of the *p*-value that you claim, so that *p*-values below this reject the null hypothesis (that is, they show that there is an effect).

Power, β level

Power is how likely you are to detect an effect if it is there. The more data you have, the more statistical power. The greater the real effect, the more power. The less random variation in your world, the more power. The more sensitive your statistical method, the more power. It

you had a method that always declared an effect significant, regardless of the data, it would have a perfect power of 1 (but it would have an α-level of 1, too, the probability of declaring significance when there was no effect). The goal in experimental design is usually to get the most power you can afford, given a certain α-level. It is not a mistake to connect the statistical term power with the common sense of power as persuasive ability. It has nothing to do with work or energy, though.

Confidence Intervals

A confidence interval is an interval around a parameter estimate that has a given probability of containing the true value of a parameter. Most often the probability is 95%. Confidence intervals are now considered one of the best ways to report results. It is expressed as a percentage of $1 - \alpha$, so an 0.05 alpha level for a two-tailed t-quantile can be used for a 95% confidence interval. (For linear estimates, it is constructed by multiplying the standard error by a t-statistic and adding and subtracting that to the estimate. If the model involves nonlinearities, then the linear estimates are just approximations, and there are better confidence intervals called *profile likelihood confidence intervals*. If you want to form confidence regions involving several parameters, it is not valid to just combine confidence limits on individual parameters.)

Biased, Unbiased

An *unbiased estimator* is one where the expected value of an estimator is the parameter being estimated. It is considered a desirable trait, but not an overwhelming one. There are cases when statisticians recommend biased estimators. The maximum likelihood estimator of the variance has a small but nonzero bias, for example.

Sample Mean versus True Mean

The *sample mean* is the one you calculate from your data—the sum divided by the number. It is a statistic, that is, a function of your data. The *true mean* is the expected value of the probability distribution that generated your data. You usually don't know the true mean, and that is why you collect data, so you can estimate the true mean with the sample mean.

Variance and Standard Deviation, Standard Error

Variance is the expected squared deviation of a random variable from its expected value. It is estimated by the sample variance. *Standard deviation* is the square root of the variance, and we prefer to report it because it is in the same units as the original random variable (or response values). The sample standard deviation is the square root of the sample variance. The term standard error describes an estimate of the standard deviation of another (unbiased) estimate.

Degrees of Freedom

The specific name for a value indexing some popular distributions of test statistics. It is called degrees of freedom because it relates to differences in numbers of parameters that are or could

be in the model. The more parameters a model has, the more freedom it has to fit the data better. The DF (degrees of freedom) for a test statistic is usually the difference in the number of parameters between two models.