# Originality in the Arts and Sciences: Lecture 2: Probability and Statistics

Let's face it. Statistics has a really bad reputation. Why?

1. It is boring.
2. It doesn't make a lot of sense.

Actually, the reason statistics is boring is that (as with everything else you think is boring) the problem is you don't really understand it, at least not on a conscious level. But in fact, all of you use statistics constantly in your evaluation of the world around you, from deciding whether it is worth the risk to ask someone on a date (do people still date?) to deciding what courses to take to get into medical school. Anytime you make a decision, you are applying statistics in some form – you probably just don't know it.

What we will do the next two days is look at why statistics is especially important to the experimental scientist and look at how we can take the subject from that rather hazy qualitative state in your head to a more quantitative playing field that everyone can understand.

But first, and interesting example of statistics in action:

## The Case For (Against?) Second Hand Smoke.

Now we all know that second hand smoke is bad for us – worse than that, it kills. And how do we know this? Because we heard it on the news, from some scientists who said that second hand smoke causes cancer.

But how did those scientists come to that conclusion? Did they put a rat in a room with 1000 burning cigarettes to see whether the rat got a spot on its lung? No. And why not? Because that experiment on a single rat wouldn't say with certainty that the rat got that spot from the smoke. But even if you used 1000 rats, people would say, who cares about rats, what about people?

Well, there are very few people who would sit in a room with 1000 cigarettes burning (unless they were smoking too which is the real reason people start to smoke) so instead, scientists had to go out and create a study with a bunch of people who hung around second hand smoke whether they wanted to or not (like spouses and children of heavy smokers.) The scientists also had to find a control group of spouses and children of people who didn't. Then after several years the scientists looked to see whether second-hand smokers were more likely to get some terrible disease than the non-second-hand smokers. And guess what? They did – well sort of. As you will see, it is impossible to be 100% CONFIDENT, at least according to the statistics. But you don't get your research funded by saying you aren't sure, so you say YES, second-hand smoke causes cancer.

In fact, you did a t-test and your statistics told you could only be about 90% CONFIDENT. In other words, for the same data, 90 times out of 100 it would be true that second-hand smoke caused cancer, but 10 times out of 100, second-hand smoke wouldn't cause cancer. Well, who decides whether being 90% CONFIDENT means SECOND HAND SMOKE CAUSES CANCER? Simple. Whoever is the boss decides. And who is the boss for the scientists who presented the study? Well, if you are trying to prove cigarette smoking causes cancer, and you want to please the NIH who gave you the money, you go on TV and say SECOND HAND SMOKE CAUSES CANCER. But if you work for a cigarette manufacturer, you go on TV and say THE RESULTS ARE INCONCLUSIVE. Who's right? You both are.

So you see why the general population hates statistics. But this story is pretty interesting, and it is an excellent reason to know something about statistics the next time you listen to someone argue a point you disagree with. Then maybe you can stand up and say, "Excuse me, what confidence limit did you use in that study?"

## Probability

As a foundation for an understanding of concepts such as sampling theory and signal to noise measurement in data collection, it is necessary to briefly discuss PROBABILITY. Probability is he tendency of an event to take place. For some event x, there is an event space with n outcomes for which x occurs m times. The probability of x occurring is

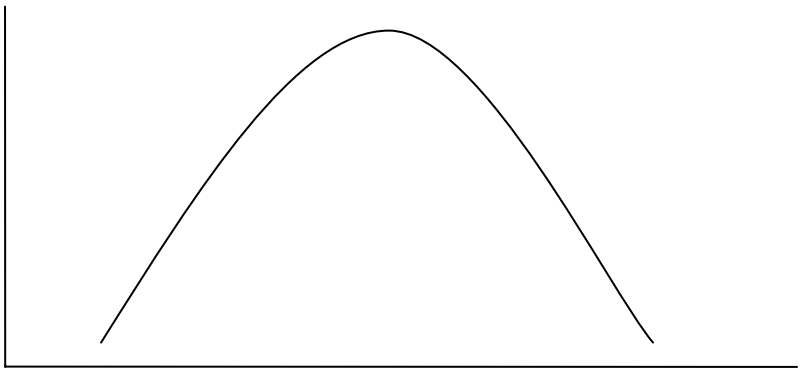
$$P(x) = m/n,\ 0 < P < 1 \qquad \text{eq 1}$$

Any probability for which all outcomes are known is objective probability. This is typically the type of probability with which we most often work. A plot of probability vs. m/n is a probability distribution which can be discrete, as with rolling a die (1,2,3,4,5,6), or continuous, as in asking how tall people are. We deal with distributions most often by describing a probability density function:

$$P(x_a < x < x_b) = \frac{\int_{x_a}^{x_b} f(x)dx}{\int_{-\infty}^{\infty} f(x)dx} \qquad \text{eq 2}$$

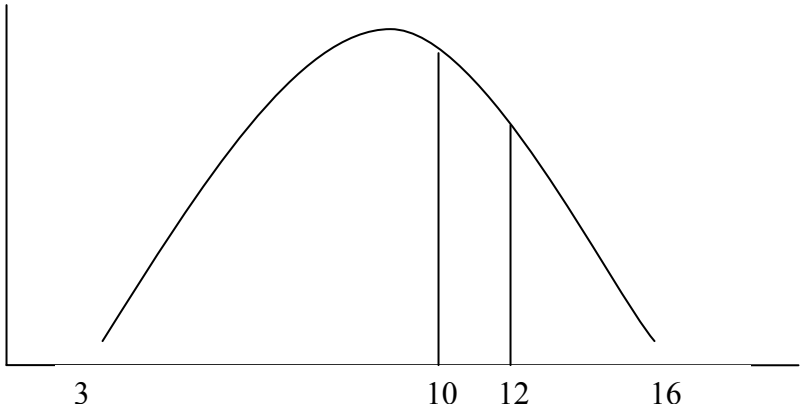
Where f(x) is the probability density function for x.

Example. Consider the shoe sizes of all the students at UT Austin – the range will be something like 3 to 16 inches. Assume that any size in this range is possible (a continuous function) and you get a distribution that




looks like the one above

If this distribution is given by a density function then the probability of finding a student with shoe size between 10 and 12 inches is given by the following area under the curve:




So what are some of the density functions we come across as scientists? Consider random variables, i.e., those results of experiments that are affected by chance. Examples include:

1. Binomial random variable. This type of distribution is found when chances are discrete, for example when flipping a coin which can give a result which is either heads or tails.
2. Geometric random variable.
3. Poisson random variable. (The observation of discrete events in a continuous interval.)
4. Uniform and exponential random variable.
5. Normal random variable (Gaussian distribution.)

**The Normal Random Variable**

Case 5 above is by far the most commonly studied because it describes the distribution of events in a continuous interval. The density function for a normal distribution is

$$f(x)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right],-\infty<x<\infty \quad \text{eq 3}$$

With two parameters

Mean = $$\mu=\int_{-\infty}^{\infty} xf(x)dx \quad \text{eq 4}$$

Variance = $$\sigma^2=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx \quad \text{eq 5}$$

Or when written as summations

$$\mu=\sum_{i=1}^{N} x_i/N \quad \text{eq 6}$$

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \quad \text{eq 7}$$

With N=number of elements considered.

Standard normal variate.

An immediate concern is how to integrate the density function above every time we have a new collection of data. Since every different distribution has a different mean and standard deviation, this could be a problem. But what if we had some standard condition, like assuming that the mean, μ, is at 0 and the standard deviation,

σ, is 1. As you can see from looking at the function, a change in mean merely shifts the position on the curve and a change in standard deviation merely expands or contracts the distribution by a constant. THE SHAPE OF THE CURVE DOES NOT CHANGE. Thus you can use a single table of data (next page) for calculations involving a Gaussian distribution. To do this, use the following transformation:

$$x = N\left(\mu,\sigma^2\right) = N(0,1)$$

$$Z_i = \left(x_i - \mu\right)/\sigma$$

*then* eq 8

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}$$

We see that this new function for the Gaussian distribution, f(Z) is the same as equation with the exception that we have applied it to the case of μ=0 and σ=1.

## POPULATIONS AND SAMPLES.

We could, of course, find the distribution of shoe sizes at UT by lining everyone up at a Payless Shoes and getting their measurement. Then we could calculate a **true mean**, μ, and a true standard deviation, σ.

You can imagine that chances of getting all the students at UT to show up at a Payless Shoes when we want is pretty unlikely. So what do we do? I guess we could get some of the students to show up, maybe by offering pizza. Then we could get a SAMPLING of the distribution of shoe sizes, with and average, χ, and a standard deviation, s

$$x = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \text{eq 9}$$

$$s^2 = \sum_{i=1}^{n} \frac{\left(x_i - x\right)^2}{\left(n-1\right)} \qquad \text{eq 10}$$

But now the big question… How much do χ and s differ from μ and σ? In other words, how much do the sampled values differ from the true value for the population?

It turns out this question can be answered by examining the distribution of experimental data. Two common sense thoughts in answering the question are that the certainty that χ and s approximate μ and σ increases (the error range decreases) when:

1. The number of samples we measure gets larger

AND

2. The distribution of s is narrow

So here in equation 11 is the BIG equation. Described mathematically, for a sample, n, of a population, N, the range of uncertainty is

$$\left(x - Z\frac{s}{\sqrt{n}}\right) \le \mu \le \left(x + Z\frac{s}{\sqrt{n}}\right) \qquad \text{eq 11}$$

Here, the uncertainty, or error range, between the sampled average, $\chi$, and the true mean, $\mu$, to achieve a certain confidence is

$$Z\frac{s}{\sqrt{n}}$$

Note this equation satisfies our two common sense ideas about how the certainty of the sampling improves as s decreases and n increases.

What is Z in equation 11? It is the distance $\chi$ is from $\mu$, given in units of standard deviation. Associated with Z is a degree of confidence (confidence level) which corresponds to the area under the Gaussian curve in the table on the previous page.

TABLE 1

| **Z (number of standard deviations)** | **Confidence level (area under the Gaussian curve)** |
|---|---|
| 1.64 | 90% |
| 1.96 | 95% |
| 2.58 | 99% |

In interpreting the table, we would say, for example, that there is a 95% confidence (95 times out of 100) that the true mean, $\mu$, lies within an interval of about +/- two (1.96) standard deviations of $\chi$.

Now remember, the values in the table are for a STANDARD NORMAL VARIATE. You have to use eq 8 to scale to the range of uncertainty for your particular data.

## SMALL SAMPLING SIZES

We can use Table 1 when we have a large enough sample size so that s approximates $\sigma$. But some times this doesn't happen if we don't collect enough data. For example, what do you do if you don't want to do 100 titrations or if you can't get 100 people to give you their shoe size?

When the number of samples is small, we can no longer use Z values in finding the area under the curve for our calculations. Rather, we substitute a t value to calculate confidence limits.

$$\left(x - t\frac{s}{\sqrt{n}}\right) \le \mu \le \left(x + t\frac{s}{\sqrt{n}}\right) \qquad \text{eq 12}$$

Where t is the number of degrees of freedom for a particular experiment. There are tables that give you this t value for a particular confidence level. For example, for a 95% confidence level:

Table 2

| Degrees of freedom | 95% confidence level |
|---|---|
| 1 | 12.71 |
| 2 | 4.30 |
| 5 | 2.57 |
| 10 | 2.23 |
| 30 | 2.04 |
| 100 | 1.98 |
| $\infty$ | 1.96 |

Notice that as the size of our sample approaches infinity, our sampled distribution, s, approaches σ, and t approaches Z. In practice, this use of t values expands the width of the interval within which we can assume a certain level of confidence. This makes sense, because if we don't have enough data to know what the true distribution looks like, we are less confident and consequently need a larger error window to speak with the same confidence.

A final point, remember that when working with small data sets, you use (n-1) instead of n, in calculating equation 12.

Now an example:

**Example:** A chemist obtained the following data for the alcohol content of a sample of blood; percent ethanol : 0.084, 0.089, and 0.079. Calculate the 95% confidence limit for the mean assuming (a) no additional knowledge about the precision of the method and (b) that on the basis of previous experiences s → σ = 0.006% ethanol.

$$\sum x_i = 0.084 + 0.089 + 0.079 = 0.252$$

(a)
$$\sum x_i^2 = 0.007056 + 0.007921 + 0.006241 = 0.021218$$

$$s = \sqrt{\frac{0.021218 - (0.252)^{2/3}}{3-1}} = 0.005$$

Here, χ = 0.252/3 = 0.084. Table 2 indicates that t=±4.30 for two degrees of freedom and 95% confidence. Thus,

$$95\%C.L. = x \pm \frac{ts}{\sqrt{n}} = 0.084 \pm \frac{4.3 \times 0.005}{\sqrt{3}} = 0.084 \pm 0.012$$

(b) Because a good value of σ is available,

$$95\%C.L. = x \pm \frac{z\sigma}{\sqrt{n}} = 0.084 \pm \frac{1.96 \times 0.006}{\sqrt{3}} = 0.084 \pm 0.007$$

Note that a sure knowledge of σ decreased the confidence interval by almost half.

## ARE THINGS THE SAME?

Often in experimental science, the question is raised as to whether two things are the same. In particular, if we have two distributions of data we can apply statistical test to answer questions like:

| | |
|---|---|
| Are two means the same? | Use the t test. |
| Are two variances the same? | Use the F test. |
| Is a piece of data bad? | Use the Q test. |

### Significance and the Null Hypothesis

Random error in systems that are free of systematic error can be tested to determine whether differences between two results are SIGNIFICANT. Significance tests are used to evaluate whether the differences are due to RANDOM variation or whether they are due to some kind of SYSTEMATIC difference (systematic or determinate error.) To perform a significance test we test the truth of a NULL HYPOTHESIS. The null hypothesis assumes that the analytical method IS NOT subject to systematic error. In other words there is no difference between the two data sets that cannot be attributed to random fluctuation. Assuming the null hypothesis is true allows a calculation of the probability that the difference between sample statistics (x and s)

and the true value (μ, σ) are from random error. The lower the probability that the observed difference occurs by chance, the less likely the null hypothesis is true. Typical probabilities used in rejecting the null hypothesis are 0.05 (5% significance or 95% confidence) and 0.01 (1% significance or 99% confidence). At the 5% level when we reject the null hypothesis, it is actually true 1 time in 20. At the 1 % level, it is actually true 1 time in 100. Thus we never prove the null hypothesis, we can only say that it has not been demonstrated to be false.

**Implementing the null hypothesis:**

1. Calculate $t_{exp}$ by rearranging equation 12 and substituting in the appropriate data to obtain:
$$\mu = x \pm \left(ts/\sqrt{n}\right)$$
eq 13
$$rearranges\ to\ t_{exp} = \left[(x-\mu)\sqrt{n}\right]/s$$
2. A $t_{table}$ value is then obtained from the appropriate distribution table for a given test (t, F, Q for example), a given significance (1%, 5% for example) and a given number of degrees of freedom (v).
3. $t_{exp}$ and $t_{table}$ are compared:

   If $t_{exp} > t_{table}$ then the null hypothesis is rejected and there is no evidence of systematic error: i.e., there is difference between the data sets which is not random.

   If $t_{exp} < t_{table}$ then the null hypothesis is not disproved and there is no evidence of systematic error; i.e., we can assume the data sets are the same.

Example of the Null Hypothesis in Action

A standard reference sample (from the NIST) contains 38.9% Hg. Three measurements on a previously calibrated spectrometer yield values of 38.9%, 37.4% and 37.1%. Is there systematic error? In other words, is the instrument broken?

Perform the three steps in the box:

1. Calculate x = 37.8% and s = 0.964%
2. From tables for a t test distribution, with n = 2 and a significance of 0.05 (5%0, $t_{table} = 4.3$
3. $t_{table} > t_{exp}$ so there is no evidence of systematic error. In other words, the instrument is not broken. The variations, 95 times out of 100, would be due to random fluctuation.

## Comparison of Distributions

Now let's apply the null hypothesis to some quantitative statistical measures for evaluating the degree of sameness—tests like the t-test, the f-test, the x2-test, Q-test, and the correlation coefficient. Each of these methods are based upon relatively simple functions, so the math involved and the procedure to implement them is trivial. For example, the first four are just variations on applications of the null hypothesis. The greater challenge is the ability to know when to apply the correct test. We will look at examples of the t and Q test which have a statistical foundation, and also the correlation function, which as commonly applied, does not.

**Statistics test example 1: The student's t-test.**

You will use the t-test when you want to know whether two distributions have the same mean. This is the most commonly used of the statistical test. It is applied in three steps:

1. First, estimate the standard error of the difference in the two means from a pooled variance.

$$s_D = \left[ \frac{\sum (x_i - x_1)^2}{N_1 + N_2 + 2} \left( \frac{1}{N_1} + \frac{1}{N_2} \right) \right]^{1/2} \qquad \text{eq 14}$$

With N1 and N2 the number of elements in population 1 and 2, $x_1$ and $x_2$ the means of the two populations.

2. Calculate $t_{exp}$ in order to apply the null hypothesis:

$$|t_{\exp}| = |(x_1 - x_2)/s_D|$$

3. Finally, we determined the significance of t for a pooled distribution of $N_1+N_2-2$ degrees of freedom, v. (P=significance, A=confidence)

$$P = 1 - A = 1 - \frac{1}{\upsilon^{1/2}\beta} \int_{-t}^{t} \left( 1 + \frac{x^2}{\upsilon} \right)^{(\upsilon+1)/2} dx \qquad \text{eq 15}$$

where β is the Beta function

$$\beta = \int_0^t t^{[(\upsilon/2)-1]} (1-t)^{-1/2} dt$$

Yikes. How do you calculate that integral? Fortunately, there are tables of these values, for a particular confidence or significance, and for a specific number of degrees of freedom.

**A few student t test table values for solution of the integral in eq 15.**

| | Sign.=.50 | Sign.=.10 | Sign.=.05 | Sign.=.01 |
|---|---|---|---|---|
| v | Conf.=5% | Conf.=10% | Conf.=95% | Conf.=99% |
| 1 | 1.00 | 6.31 | 12.71 | 63.7 |
| 4 | .74 | 2.13 | 2.80 | 4.60 |
| 7 | .71 | 1.9 | 2.37 | 3.50 |
| 20 | .69 | 1.72 | 2.09 | 2.84 |
| ∞ | .67 | 1.65 | 1.96 | 2.58 |

Example of t-Test

$^{14}C0_2$ is often used as a tracer for plant metabolism. You feed the radioactive $^{14}C0_2$ to the plant, let the plant use it to produce metabolic products, and then apply an analytical technique that measures radioactivity to determine whether a particular compound isolated from the plant has incorporated the labeled $^{14}C0_2$.
In one test a compound isolated from the plant gives radioactivity counts of 28, 32, 27, 39 and 40 counts/minute. A blank run of unlabeled $C0_2$ yields background counts of 28, 21, 28 and 20 counts/minute. (Remember, there are cosmic rays flying through the air create this noisy background.) Can we be confident that the compound has incorporated the labeled $^{14}C0_2$? Consider for both 95% and 99% confidence levels.

$$Calculate: x_1 = 33.2, x_2 = 24.2, s_D = 3.6$$

$$|t_{exp}| = \left(\frac{33.2 - 24.2}{3.6}\right) = 2.5$$

Now look at the tables for the case of n=7 degrees of freedom.
For 95% confidence, $t_{table}$=2.37, so $t_{exp} > t_{table}$
For 99% confidence, $t_{table}$=3.70, so $t_{exp} < t_{table}$

Thus we can say with 95% confidence that the means are different and that there is labeled $^{14}C0_2$ in the extracted compound. But we cannot say this with 99% confidence!!

**Statistics Test Example 2: The Q-Test for Rejecting Bad Data**

Often we encounter data which appears to be SIGNIFICANTLY different from the rest of the data set due to some type of systematic error. We can evaluate statistically using the null hypothesis whether we have any statistical validity in throwing out data by using the Q test.

The test is simple to implement:

1. Line up an array of data.
2. Calculate the range for the data (the difference between the largest and smallest value in the array).
3. Calculate the gap (the difference between the questionable data and its nearest neighbor.
4. Calculate $Q_{exp}$ = gap/range.
5. Compare to $Q_{table}$.
6. Apply the null hypothesis (throw out data if $Q_{exp} > Q_{table}$).

**Partial Q table**

| No. of data | 10% significance 90% confidence | 5% significance 95% confidence | 1% significance 99% confidence |
|---|---|---|---|
| 3 | 0.94 | .97 | .99 |
| 4 | .76 | .83 | .93 |
| 5 | .64 | .71 | .82 |
| 6 | .56 | .63 | .74 |
| 7 | .51 | .57 | .68 |
| 8 | .47 | .53 | .63 |
| 9 | .44 | .49 | .60 |

Example of Q Test:

You've received tests averages in my class of 77, 85, 88, 89, and 93. You come in to argue that the 77 is an aberration and should be thrown out when I assign a grade. Do you have a statistical leg to stand on?

Calculate: Range=16 Gap=8 $-Q_{exp}=8/16=0.5$

$Q_{table}$= from 0.64 (at 10% significance) up to 0.82 (99%) for 5 data points. In each case, $Q_{exp}<Q_{table}$ which means I must keep the score.

## LINEAR CORRELATION

How often do you use the word CORRELATE? Actually it is a term we throw around all the time when we want to act like we know what we are talking about. But what is its mathematical meaning? There are several, but to most of you it is that r button on your calculator. It has to do with how well a pair of quantities ($x_i$, $y_i$), I=1,2,…,N are associated. You know that:

r=1 means there is a positive correlation between x and y
r=-1 means there is a negative correlation between x and y
r=0 means there is no correlation between x and y

But how is the calculation done? Ways to calculate the correlation coefficient, r, are numerous, but often used is the Pearson coefficient:

$$r = \frac{\sum (x_i - x)(y_i - y)}{\left(\sum_i (x_i - x)^2\right)^{1/2}\left(\sum_i (y_i - y)^2\right)^{1/2}} \qquad \text{eq 16}$$

**Example of Correlation Coefficient calculation**

You are trying to match up the spectrum of an unknown compound with about a million known spectra stored in a spectral library on a computer. You need to give the computer a way to do the calculation. The correlation coefficient is one possibility, because when there is a good match of the two spectra, r should approach2.

| Mass | Library Intensity | Unknown Intensity |
|---|---|---|
| 28 | 24 | 17 |
| 29 | 17 | 14 |
| 43 | 97 | 100 |
| 57 | 100 | 99 |
| 58 | 3 | 0 |
| 75 | 17 | 8 |
| 101 | 16 | 17 |
| 103 | 5 | 0 |

From eq 18

$$x = 34.87, y = 31.87$$

$$\sum (x_i - x)(y_i - y) = 11359$$

$$\left(\sum (x_i - x)^2\right)^{1/2} = 101.8$$

$$\left(\sum (y_i - y)^2\right)^{1/2} = 111.9$$

$$r = 11359/(101.8)(111.9) = 0.997$$

This number indicates a strong degree of correlation between the two spectra and consequently you might be correct in assuming the unknown is identified as the library spectrum.
BUT
Notice that there is NO statistical validity to the degree of correlation between the correlation. (In other words, there is no standard deviation used in the calculation of eq 18.) Why does this matter? Well, there are plenty of ways that skew the data to obtain an artificially large r value. You will do this in your homework.

In fact, there are better ways to incorporate standard deviation into the uncertainty associated with r, but most people assume s=1 just to make the calculation easier (for example, your calculator probably calculates r with the eq 18 assuming x=1.) So the next time you are in a class with a professor throwing around r = 0.99995 values and saying there is a good correlation, ask about the statistics.