Revised: February 11, 2010

Notes on Statistics

Lecture I

Carl W. Akerlof

"There are three kinds of lies: lies, damned lies, and statistics." – Mark Twain (Autobiography – 1924)

Why Statistics?

1. Many important discoveries are made at the edge of technical feasibility. Every little bit of improvement in the signal-to-noise ratio helps. Example: Nobel Prize in 2006 to John C. Mather and George F. Smoot for measurements of the cosmic microwave background anisotropy.
2. Often the results of experiments are numerical values used to make further predictions. One would like to determine such numbers as accurately as possible.

My own occasional efforts in this area:

1. "Application of cubic splines to the spectral analysis of unequally spaced data", Astrophysical Journal **436**, 787-794 (1994). (method for finding periodically variable stars).
2. "Astronomical image subtraction by cross-convolution", Astronomical Journal **677**, 808-872 (2008). (method for finding optical transients such as supernovae)

Notation for expectation values:

$$\overline{x} = \langle x \rangle = E(x)$$

Measurements:

Either discrete (like counting events or people) or continuous (like measuring a length). For discrete, associate a probability, p_i, for the $i'th$ value.

$$\sum_{i=1}^{n} p_i = 1$$

For continua, associate a probability distribution function (PDF), $f(x)$, so that the probability of observing a value on the interval $[x,\ x+dx]$ is given by $f(x)dx$.

$$\int_{x_{\min}}^{x_{\max}} f(x)dx = 1$$

Probability distributions come in all shapes and flavors although the most important is the Gaussian:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(x-\overline{x})^2}{2\sigma^2}}$$

where $\overline{x}$ is the mean and σ^2 is the variance. The Gaussian has some very nice mathematical properties: it is defined everywhere from $-\infty$ to $+\infty$ and it is differentiable to all orders. Unfortunately, its integral is a transcendental function that must be evaluated by suitable approximations. In Excel, use ERF or ERFC but divide the argument by $\sqrt{2}$.

The mean of a continuous distribution is defined by

$$\overline{x} = \langle x \rangle = \int_{x_{\min}}^{x_{\max}} x f(x) dx$$

The variance is defined by

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2 = \int_{x_{\min}}^{x_{\max}} (x - \overline{x})^2 f(x) dx$$

These two moments are the most important quantities for characterizing a probability distribution. Two higher order moments are sometimes computed, skewness and kurtosis, but they have little use in physics or astronomy.

For discrete variables, the mean and variance are defined by the analogous summations:

$$\overline{x} = \sum_i x_i p_i$$
$$\sigma_x^2 = \sum_i (x_i - \overline{x})^2 p_i$$

One of the simplest probability distribution functions is the uniform distribution:

$$f_u(x) = \begin{cases} 1 & 0 \le \mathrm{x} \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

This distribution is particularly easy to synthesize with a digital computer and is the root of almost every computer simulation of probabilistic phenomena.

Problem 1: Find the mean and variance for the uniform probability distribution defined above.

We now ask: what is the distribution of the sum of the two quantities, each uniformly distributed, $x_{\Sigma 2} = x_a + x_b$? The result is obtained by convolution:

$$f_{\Sigma 2}(x)=\begin{cases}\int_0^x dt = x; \quad 0\le x\le 1\\ \int_{x-1}^1 dt = 2-x; \quad 1\le x\le 2\\ 0 \ \text{ otherwise}\end{cases}$$

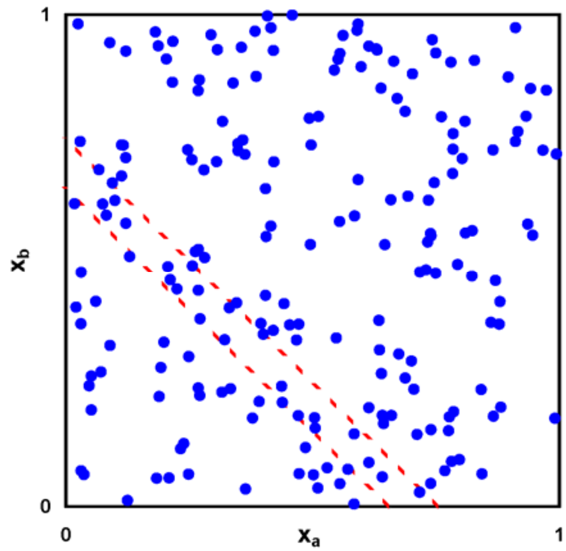

Figure 1. The geometry corresponding to the addition of two independent uniformly distributed variables. In the region bounded by the two dotted red lines, the sum, $x_a + x_b$, is approximately constant. The sum can range from 0 to 2 with a maximum of 1.

The results for larger sums of uniformly distributed numbers can be obtained by iteration:

Distribution for sum of 3 uniformly distributed values:

$$f_{\Sigma 3}(x)=\int f_{\Sigma 2}(x-t)f_u(t)dt$$

$$=\begin{cases}\frac{1}{2}x^2; \quad 0\le x\le 1\\ \frac{1}{2}(-3+6x-2x^2); \quad 1\le x\le 2\\ \frac{1}{2}(9-6x+x^2); \quad 2\le x\le 3\\ 0 \ \text{ otherwise}\end{cases}$$

Distribution for sum of 4 uniformly distributed values:

$$f_{\Sigma 4}(x) = \int f_{\Sigma 3}(x-t) f_u(t)dt$$

$$= \begin{cases} \frac{1}{6}x^3; \;\; 0 \le x \le 1 \\ \frac{1}{6}(4 - 12x + 12x^2 - 3x^3); \; 1 \le x \le 2 \\ \frac{1}{6}(-44 + 60x - 24x^2 + 3x^3); \;\; 2 \le x \le 3 \\ \frac{1}{6}(4-x)^3; \;\; 3 \le x \le 4 \\ 0 \;\; \text{otherwise} \end{cases}$$

These four distributions for the sum of 1 to 4 uniformly distributed values are plotted in figure 2 below. In addition, the Gaussian distributions with identical mean and variance for $\Sigma 3$ and $\Sigma 4$ are shown as dotted lines. I hope the trend is clear: the sum of a set of uniformly distributed numbers approaches a Gaussian distribution ever more closely as the number of items increases. In fact, a good computer algorithm for generating Gaussian random numbers simply adds twelve uniformly distributed numbers on the interval, [0,1], and subtracts the mean of 6.

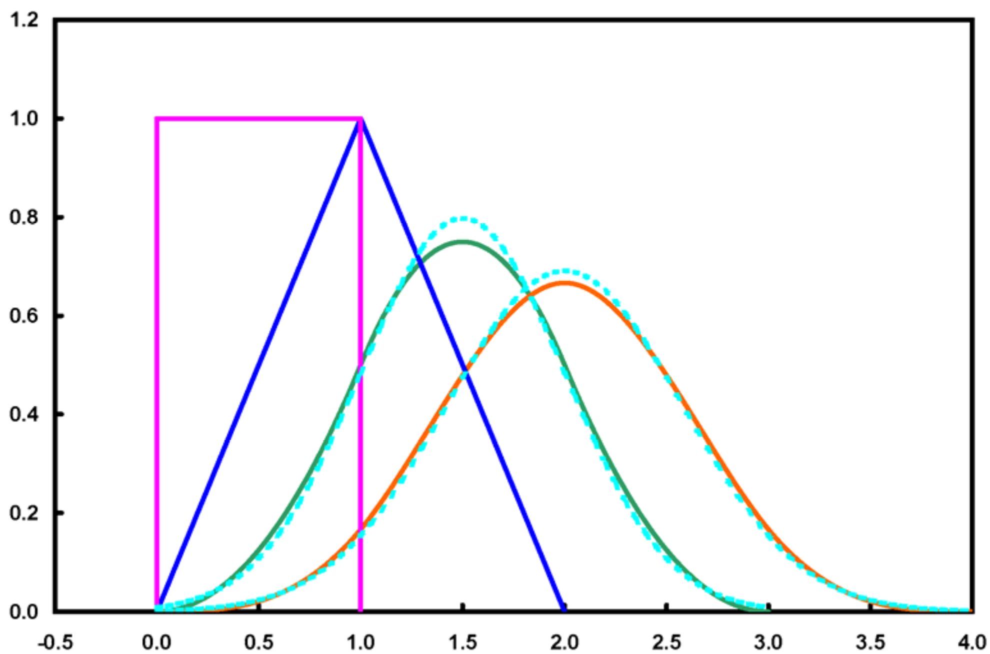

Figure 2. Plots of the distribution function for the sum of the one to four uniformly distributed numbers. Magenta = 1, blue = 2, green = 3, orange = 4. The dotted lines show the Gaussian distributions corresponding to the mean and variance for 3 and 4 summed values.

Problem 2: Compute the mean and variance for the $f_{\Sigma 2}$, $f_{\Sigma 3}$ and $f_{\Sigma 4}$ distributions. Suggestion: Use Mathematica or a similar program to perform the integrations symbolically.

The results shown above are quite general. Under very broad limits, any sum of randomly distributed variables approach Gaussian distributions as the number of variables as increases. This result is called the Central Limit Theorem and gives some idea of why the Gaussian distribution plays such a significant role in statistics.

More generally, the following relationships for the mean and variance are independent of the underlying distribution functions:

$$\overline{a+b} = \langle a+b \rangle = \int\int (t_a + t_b) f_a(t_a) f_a(t_b) dt_a dt_b$$

$$= \int (\langle a \rangle + t_b) f_b(t_b) dt_b$$

$$= \langle a \rangle + \langle b \rangle$$

$$\sigma_{a+b}^2 = \langle (a+b)^2 \rangle - \langle a+b \rangle^2 \int\int (t_a + t_b - \langle a \rangle - \langle b \rangle)^2 f_a(t_a) f_b(t_b) dt_a dt_b$$

$$= \int (\langle a^2 \rangle - \langle a \rangle^2 + \langle b \rangle^2 - 2\langle b \rangle t_b + t_b^2) f_b(t_b) dt_b$$

$$= \langle a^2 \rangle - \langle a \rangle^2 + \langle b^2 \rangle - \langle b \rangle^2 = \sigma_a^2 + \sigma_b^2$$

These relations can be extended to an arbitrary number of variables, independent of the distribution function for each element.

Scaling the mean and variance:

If $g(x) = cx$ where c is a constant, then $\overline{g} = c\overline{x}$; $\sigma_g^2 = c^2 \sigma_x^2$

If $g(x)$ is a more complicated function, reasonable approximations are given by:

$$\overline{g} \cong g(\overline{x}); \qquad \sigma_g^2 \cong \left(\left. \frac{dg}{dx} \right|_{x=\overline{x}} \right)^2 \sigma_x^2$$

as long as higher-order derivatives are not important over the range, $[x - \sigma_x,\ x + \sigma_x]$. If this becomes questionable, it is best to estimate the statistical parameters by Monte Carlo simulation.

Note that I have carefully avoided discussing the standard deviation which is defined as the square root of the variance. Although the standard deviation is important, it is the variance that is the more fundamental statistical quantity.

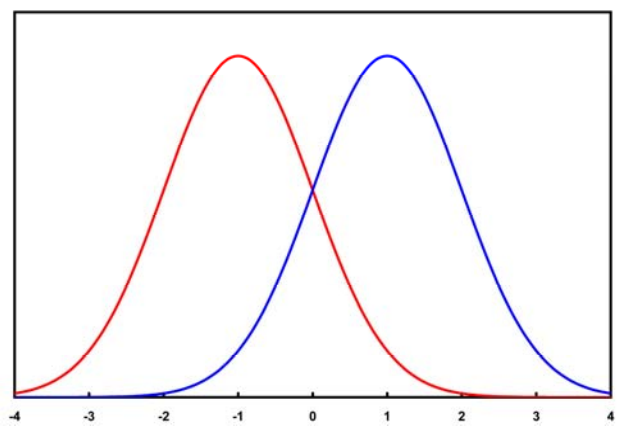

Figure 3. Two distributions with different means but identical variances (standard deviations).

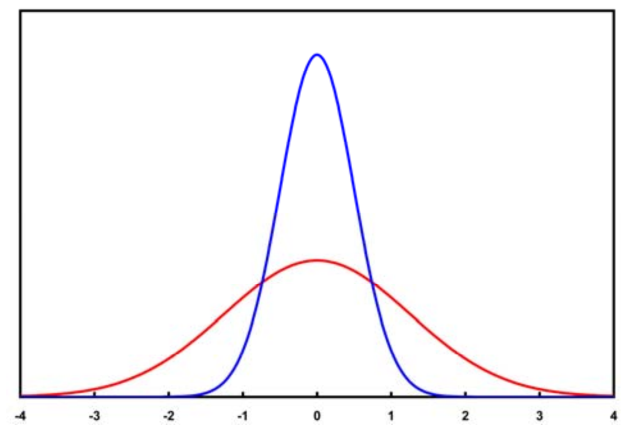

Figure 4. Two distributions with identical means but different variances (standard deviations).

Binomial distribution:

There are two common discrete distributions that one usually encounters in science. One of these is the binomial distribution that describes the outcome of such activities as coin tossing. Suppose, for the sake of argument, that coins have probability p of landing "heads" and a probability q of landing "tails" $(q = 1 - p)$, If the coins are "fair", $p = q = 0.5$, but that may not always be true. Suppose we toss N coins. The probability of all possible outcomes is given by:

$$(p+q)^N = 1$$

The expression on the left side of this equation can be rewritten as:

$$(p+q)^N = \sum_{m=0}^{N} \frac{N!}{m!(N-m)!} p^m q^{N-m}$$

Each individual term in this sum represents a different outcome of the total number of heads and tails with m denoting the number of "heads" and N-m the number of "tails". Therefore, the binomial distribution function for the probability of observing each value of our random variable, m, is

$$P(m) = \frac{N!}{m!(N-m)!} p^m q^{N-m}$$

Note that the total number of heads (or tails) is strictly bound between the limits of 0 and N. This is distinctly different from the Poisson distribution that will be discussed shortly where no such upper bound exists.

It is fairly straight-forward to show that the mean and variance are given by:

Mean: $\bar{n} = p\ N$

Variance: $\sigma_n^2 = p\ q\ N$

Proof:

$$\bar{n} = \langle n \rangle = pN \sum_{m=1}^{N} \frac{(N-1)!}{(m-1)!(N-m)!} p^{m-1} q^{N-m}$$
$$= pN \sum_{m=0}^{N-1} \frac{(N-1)!}{m!(N-1-m)!} p^{m} q^{N-1-m} = pN$$

$$\langle n^2 \rangle - \langle n \rangle = p^2 N(N-1) \sum_{m=2}^{N} \frac{(N-2)!}{(m-2)!(N-m)!} p^{m-2} q^{N-M}$$
$$= p^2 N(N-1) \sum_{m=0}^{N-2} \frac{(N-2)!}{m!(N-2-m)!} p^{m} q^{N-2-m} = p^2 N(N-1)$$

thus $\sigma_n^2 = \langle n^2 \rangle - \langle n \rangle^2 = p^2 N(N-1) + \langle n \rangle - \langle n \rangle^2$
$= p\ q\ N$

Note that both the mean and the variance scale directly with the number of events, *N*. Thus, the precision of an observation, given by the ratio of the standard deviation to the mean decreases at a rate proportional to $1/\sqrt{N}$. Making *N* as large as possible is good but you improve more and more slowly for the extra time and resources required.

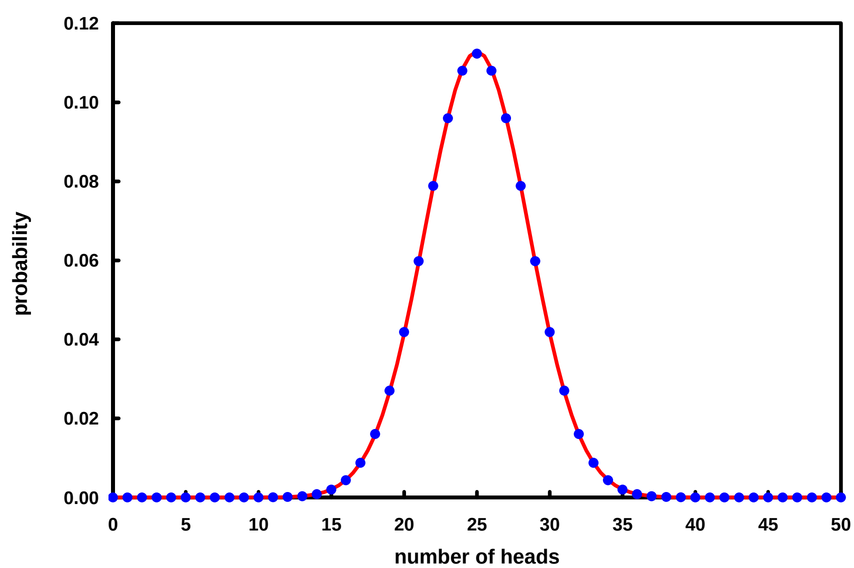

Figure 5. Probability distribution for the number of "heads" observed when tossing 50 coins. The points show the exact probability computed from the binomial distribution. The smooth curve is the Gaussian distribution with $\bar{n}$ = *25* and σ_n^2 = *12.5*. At the peak of the distributions, the two probabilities differ by 0.5%.

Figure 5 shows a plot of these terms for $N = 50$ for "fair" coins. As expected from the Central Limit Theorem, the points follow very closely a Gaussian distribution with the mean and variance computed from the formulas above with $p = q = 0.5$.

Poisson distribution:

The most important statistical distribution function for Physics 441/442 is the Poisson – it describes the probability of finding *n* discrete events in a specific time interval assuming uniform arrival probability. Unlike the binomial distribution, the number of events is not fixed. In principle, a very large number of events can occur randomly within a short interval although the probability becomes vanishingly small as *n* gets large.

Assume the average interval between events is given by:

$$\langle t\rangle = T/N_T$$

where T is the length of an appropriately long time interval and N_T is the number of corresponding events.

The probability of zero events within a time interval of length t is:

$$p(0) = e^{-t/\langle t\rangle}$$

For very short time intervals, $p(0)$ is close to one but decreases rapidly as t gets appreciably larger than $\langle t\rangle$.

For *n* events, the corresponding probability is:

$$p(n) = \frac{1}{n!}\left(t/\langle t\rangle\right)^n e^{-t/\langle t\rangle}$$

It is trivial to show that $p(n)$ reaches a maximum when $t = n\langle t\rangle$. Figure 6 below shows such behavior graphically.

One can also see that:

$$\sum_{n=0}^{\infty} p(n) = 1$$

independent of the duration, t.

The mean number of events can be computed from:

$$\overline{n} = \sum_{n=0}^{\infty} n\, p(n) = x\sum_{n=1}^{\infty}\frac{x^{n-1}}{(n-1)!}\, e^{-x} = x = t/\langle t\rangle$$

Problem 3: Using similar techniques, find the variance of n, $\sigma^2 = \langle n^2\rangle - \langle n\rangle^2$

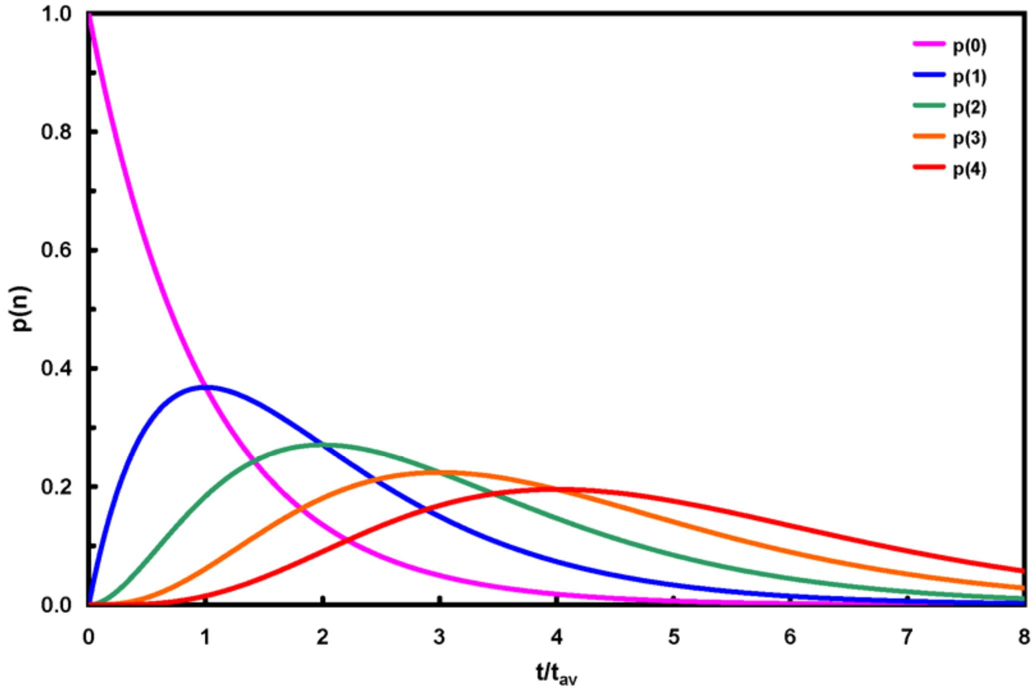

Figure 6. The probability of detecting exactly 0, 1, 2, 3, or 4 events as a function of the ratio of t divided by the average time between events, t_{av}. Note that each curve peaks when $t = n\ t_{av}$.

Appendix A

Distribution functions for sums of three and four uniformly distributed variables.

3 variables: $\int_0^1 f_2(x-t)f_1(t)dt$

$$f_3(x)=\int_0^x (x-t)dt;\quad 0\le x\le 1$$

$$f_3(x)=\int_0^{x-1} (2-(x-t))dt+\int_{x-1}^1 (x-t)dt;\quad 1\le x\le 2$$

$$f_3(x)=\int_{x-2}^1 (2-(x-t))dt;\quad 2\le x\le 3$$

4 variables: $\int_0^1 f_3(x-t)f_1(t)dt$

$$f_4(x)=\int_0^x \frac{1}{2}(x-t)^2\,dt;\quad 0\le x\le 1$$

$$f_4(x)=\int_0^{x-1} \frac{1}{2}(-3+6(x-t)-2(x-t)^2)+\int_{x-1}^1 \frac{1}{2}(x-t)^2\,dt;\quad 1\le \mathrm{x}\le 2$$

$$f_4(x)=\int_0^{x-2} \frac{1}{2}(9-6(x-t)+(x-t)^2)dt+\int_{x-2}^1 \frac{1}{2}(-3+6(x-t)-2(x-t)^2\,dt;\; 2\le \mathrm{x}\le 3$$

$$f_4(x)=\int_{x-3}^1 \frac{1}{2}(9-6(x-t)+(x-t)^2)dt;\quad 3\le x\le 4$$

Notes on Statistics

Lecture II

Carl W. Akerlof

Sample statistics with n observations:

Sample Mean: $\bar{x}_s = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample Variance: $\sigma_s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$

Why the factor of 1/(n - 1) instead of 1/n ?

$$\sigma_s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 + \sum_{\substack{i,j\\ i\neq j}} x_i x_j\right)\right)$$

$$= \frac{1}{n-1}\left(\frac{n-1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{\substack{i,j\\ i\neq j}} x_i x_j\right)$$

Since there are n terms in the first sum and $n\cdot(n-1)$ terms in the second, the average value for an ensemble of samples coincides with σ_x^2.

Sample Standard Deviation: $\sigma_s = \sqrt{\sigma_s^2}$

With these definitions, the sample statistics cluster around the population statistics:

Mean of Sample Means: $\langle \bar{x} \rangle = \langle x \rangle$

Mean of Sample Variances: $\langle \sigma_s^2 \rangle = \sigma_x^2$

The next question is how accurately do the sample statistics, $\bar{x}_s$ and σ_s^2, compare with the population statistics, $\bar{x}$ and σ_x^2? (Beware that the statistical quantities described below are population statistics based on the ensemble of all possible similar experiments, usually an infinite set.)

Variance of Sample Means: $\sigma_{\bar{x}}^2 = \frac{1}{n}\sigma_x^2$

Proof:

$$\sigma_{\bar{x}_s}^2 = \left\langle \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \right\rangle - \langle x \rangle^2$$

$$= \frac{1}{n^2} \left\langle \sum_{i=1}^{n} x_i^2 \right\rangle + \frac{1}{n^2} \left\langle \sum_{\substack{i,j \\ i \neq j}} x_i x_j \right\rangle - \langle x \rangle^2$$

$$= \frac{1}{n} \langle x^2 \rangle + \frac{n \cdot (n-1)}{n^2} \langle x \rangle^2 - \langle x \rangle^2 = \frac{1}{n} \langle x^2 \rangle - \frac{1}{n} \langle x \rangle^2$$

$$= \frac{1}{n} \sigma_x^2$$

Variance of Sample Variances:

$$\sigma_{\sigma_s^2}^2 = \frac{1}{n} \langle x^4 \rangle - \frac{4}{n} \langle x^3 \rangle \langle x \rangle - \frac{n-3}{n(n-1)} \langle x^2 \rangle^2 + \frac{4(2n-3)}{n(n-1)} \langle x^2 \rangle \langle x \rangle^2 - \frac{2(2n-3)}{n(n-1)} \langle x \rangle^4$$

With the presence of the third and fourth moments, the variance of the sample variance depends on the functional form of the probability distribution function. For the important case of the Gaussian distribution;

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$$

we can choose a coordinate system so that $\bar{x}$ is zero, $\langle x^4 \rangle = 3\sigma^4$, leading to:

Variance of sample variances: $\sigma_{\sigma_s^2}^2 = \dfrac{2\sigma^4}{n-1}$

and the related equations:

Standard deviation of sample variance: $\sigma_{\sigma_s^2} = \sqrt{\dfrac{2}{n-1}}\ \sigma^2$

Variance of sample standard deviations: $\sigma_{\sigma_s}^2 = \dfrac{\sigma^2}{2(n-1)}$

Standard deviation of sample standard deviations: $\sigma_{\sigma_s} = \dfrac{\sigma}{\sqrt{2(n-1)}}$

For the binomial distribution:

Variance of sample variances: $\sigma^2_{\sigma_s^2} = \dfrac{2\sigma_n^4}{m-1} + \dfrac{(1-6pq)\sigma_n^2}{m}$

and for the Poisson distribution:

Variance of sample variances: $\sigma^2_{\sigma_s^2} = \dfrac{2\sigma_n^4}{m-1} + \dfrac{\sigma_n^2}{m}$

where m is the number of observations. Note the similarities in form to the corresponding expression for the Gaussian distribution.

For many applications, statistical distributions are plagued by outliers that can badly contaminate the mean and variance. Statistical parameters based on the ordering of values are much more stable against the effects of poorly understood background effects. The median of a set of values is easily defined but it does require that the set first be ordered.

Variance of median: $\sigma^2_{x_{1/2}} = \dfrac{1}{4nf^2\left(x_{1/2}\right)}$; $F(x_{1/2}) \equiv \dfrac{1}{2}$

For Gaussian distributions: $\sigma^2_{x_{1/2}} = \dfrac{\pi}{2n}\sigma^2$

There are two disadvantages incurred by using the median instead of the mean to represent the average of a sample. Ordering takes a significant amount of computational effort, especially if the set is large. Secondly, for a fixed value of n, the variance of the median is $\pi/2$ greater than for the mean. (This ratio is called the Asymptotic Relative Efficiency, *ARE*.) The choice between the two will depend on the application. (There is no free lunch.)

The variance is even more susceptible to contamination than the mean. As an example, a 1% admixture of values that are 10 times larger than the σ for the rest of the sample will double the variance. Understanding background outliers at the 1% level is usually an extremely daunting task. An order statistic that will robustly estimate dispersion is the Inter-Quartile Difference (IQD), defined by:

Assume $F(x_{1/4}) \equiv \dfrac{1}{4}$; $F(x_{3/4}) \equiv \dfrac{3}{4}$

Then: $\text{IQD} = x_{3/4} - x_{1/4}$

For a Gaussian distribution:

$$\frac{\text{IQD}}{q^*} = \sigma \quad \text{where} \quad q^* = 1.3489798 \ ; \ \frac{1}{2} = \frac{1}{\sqrt{2\pi}} \int_{-q^*/2}^{+q^*/2} e^{-\frac{x^2}{2}} dx$$

$$\sigma^2_{(\text{IQD}/q^*)^2} = \frac{2\pi}{n} \frac{e^{\left(\frac{q^*}{2}\right)^2}}{q^{*2}} \sigma^4$$

$$= 2.7209 \frac{2\sigma^4}{n}$$

Note that the variance of the IQD-estimated dispersion is a factor of 2.72 larger than for the normal variance, again an example of the cost in precision that must be paid for a more reliable statistical measure.

Statistical significance:

A principle goal of statistics is to determine the probability that two measured values are consistent in view of the estimated errors. In the first lecture, several cases were shown in which the effective probability distribution functions were, to a very good approximation, Gaussian. In such case, we are interested in the cumulative distribution function:

$$F(x') = \int_{x_{\min}}^{x'} f(x) dx$$

which for a Gaussian distribution is:

$$F(x') = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x'} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx$$

In Excel, this can be easily computed from the normalized integral:

$$\text{NORMSDIST}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$$

where $u = \dfrac{x - \bar{x}}{\sigma}$. Similar functions are available on other mathematical evaluation packages.

Most of the time, we are interested in fairly simple questions such as whether a particular measurement is consistent with a previously well-established value. Suppose that you found a value for Planck's constant of 4.15×10^{-34} J-s with a standard deviation of 1.16×10^{-34} for the measurement error. The presently accepted value is 6.626069×10^{-34}. Your result is certainly numerically different. Should you call the science editor of the New York Times and inform him that a fundamental constant of physics is changing with time or perhaps location? The question that statistics can answer is the probability that such a measurement could occur by chance alone.

The normalized error in this ease is $(4.150-6.626)/1.16=-2.13$ standard deviations. The probability that a measurement would have such a departure from the true value is

$$F(-2.13)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-2.13} e^{-\frac{x^2}{2}}\,dx \simeq 0.0164$$

Given that there are 35 students in Physics 441 this term, there is greater than 50% probability that somebody will get such a large discrepancy during the next few months. We have also low-balled the likelihood of chance error: you might be equally willing to call the New York Times if the error were correspondingly higher, ie. 9.10×10^{-34}. In this case, the chance probability is

$$p\left(|x|>2.13\sigma\right)=\frac{1}{\sqrt{2\pi}}\left\{\int_{-\infty}^{-2.13} e^{-\frac{x^2}{2}}\,dx+\int_{+2.13}^{\infty} e^{\frac{-x^2}{2}}\,dx\right\}\simeq 0.0328$$

Note that statistics has no way of estimating the effects of systematic problems with your results.

Problem 4: You have made two measurements of the counting rate produced by a radioactive isotope. For the first observation, you detected 820 counts in 1000 seconds; in the second, you detected 1625 counts in 1700 seconds. Find the (two-sided) probability that the apparent difference in rates is due to chance alone.

Problem 5: You are given a single penny. How many times would you have to flip it to determine with 99% confidence that probability of heads lies in the range, $0.495 \le p(H) \le 0.505$, assuming that for your specific experiment exactly half of the tosses were "heads"?

There has recently arisen a new catechism for corporations called "6 σ". Originating with some folks at Motorola who wanted to reduce the number of defective products, the idea is to make sure that your production techniques are good enough so that only a 6 σ error would lead to failure. The true doctrine is a bit of a fake - it's actually 4.5 σ errors that are bad news and the probability is calculated as one-sided, not two.

For many cases you are likely to deal with, you may need to use the exact distribution function if the deviation from Gaussianity is large or instead rely on numerical simulations. The more anomalous or significant your result, the more care you need to take to properly calculate the statistical significance.

Weighted means:

In earlier discussion, a formula for the sample mean was presented that assumed that each of the n observations is associated with identical standard deviation (monoscedastic). That situation is frequently violated, leading to the need to find a way of weighting the formula to provide a value for the mean with the least error.

Assume a set of weights, $\{w_i\}$ so that:

$$\bar{x}_s = \sum_{i=1}^{n} w_i x_i \Big/ \sum_{i=1}^{n} w_i$$

$$\sigma_{\bar{x}_s}^2 = \sum_{i=1}^{n} w_i^2 \sigma_{x_i}^2 \Big/ \left(\sum_{i=1}^{n} w_i \right)^2$$

We must find a set of values for $\{w_i\}$ to minimize $\sigma_{\bar{x}_s}^2$:

$$\frac{\partial \sigma_{\bar{x}}^2}{\partial w_i} = \left(2 w_i \sigma_{x_i}^2 \sum_{i=1}^{n} w_i - 2 \sum_{i=1}^{n} w_i^2 \sigma_{x_i}^2 \right) \Big/ \left(\sum_{i=1}^{n} w_i \right)^3 = 0$$

This can be solved by setting $w_i = \dfrac{1}{\sigma_{x_i}^2}$ for all i.

The variance of the mean is thus:

$$\sigma_{\bar{x}_s}^2 = 1 \Big/ \sum_{i=1}^{n} \frac{1}{\sigma_{x_i}^2}$$

The formula given above for weighting observations should be followed whenever considering data with different measurement errors.

Problem 6: Find the weighted mean of the two counting rates described in Problem 4.

Notes on Statistics

Lecture III

Carl W. Akerlof

Parameter estimation:

In the last lecture, a procedure was developed for finding the mean of a set of observations with different errors. It was shown that the most accurate estimate was obtained by a weighted sum:

$$\bar{x} = \frac{\sum w_i x}{\sum w_i}$$

where:

$$w_i = \frac{1}{\sigma_i^2}$$

and thus:

$$\sigma_{\bar{x}}^2 = \frac{1}{\sum \frac{1}{\sigma_i^2}}$$

We will now take a slightly different point of view to derive the same result. This new method can be extended to much more complex descriptions of data than just the mean.

For the moment, make the assumption that every observation, if not perturbed by a variety of unwanted phenomena, would yield exactly the same value, x_m. We would like to find a value for x_m so that the difference between observed values, $\{x_i\}$, and x_m are as small as possible. The simplest function suitable for this purpose is the sum of squares of deviations, divided by the variance for each point:

$$\Delta = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(x_i - x_m \right)^2$$

This can be easily recognized as the sum of the squares of the departure of each observation from x_m in units of the standard deviation. (This expression can also be derived more formally by the maximum likelihood principle.)

Minimizing Δ by varying x_m yields the following equation:

$$\sum \frac{1}{\sigma_i^2} x_m = \sum \frac{1}{\sigma_i^2} x_i$$

or:

$$x_m = \frac{\sum \frac{1}{\sigma_i^2} x_i}{\sum \frac{1}{\sigma_i^2}}$$

as derived previously. For reasons that will soon become clear, it is useful to think of this as a simple matrix equation of the form:

$$\underline{\underline{H}}\ \underline{p} = \underline{q}$$

where $\underline{\underline{H}} = \left[\sum \frac{1}{\sigma_i^2} \right]$; $\underline{p} = [x_m]$; $\underline{q} = \left[\sum \frac{1}{\sigma_i^2} x_i \right]$

Note that for this example,

$$\sigma_{x_m}^2 = H_{11}^{-1} = \frac{1}{\sum \frac{1}{\sigma_i^2}}$$

Now assume a slightly more complicated problem, depicted in Figure 7 below. Suppose we have a total of *n* observations, half clustered around x_a and half around x_b. We would like to estimate the error of the slope of the line that best fits the entire set of data, assuming an average error of σ for the *y* values. The mean value for each of the two clusters will be known with a variance of $\frac{2}{n}\sigma^2$ and thus the variance of the slope will be determined to an accuracy of $\frac{4}{n} \frac{\sigma^2}{\left(x_b - x_a\right)^2}$.

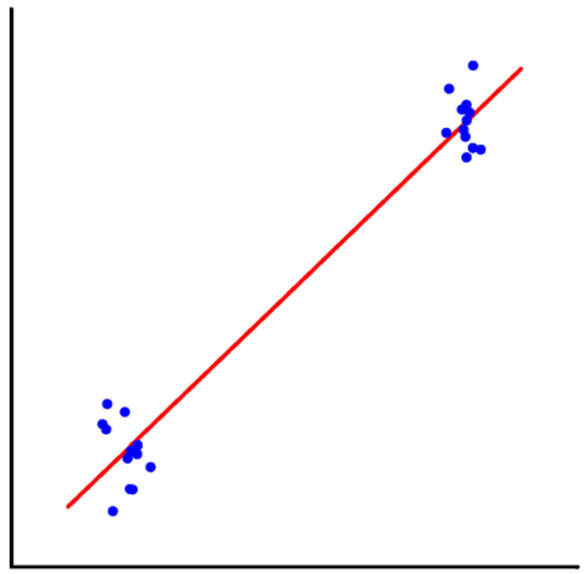

Figure 7. Least square fit of a straight line to two clusters of data points.

This problem can be framed more generally by the assumption that the data is described by a linear relationship, $y_i = A + Bx_i$. Our task is to find the best values of A and B and their variances, σ_A^2 and σ_B^2.

Using the previous idea of minimizing the square error, we have:

$$\Delta = \sum \frac{1}{\sigma_i^2}\left(y_i - \left(A + Bx_i\right)\right)^2$$

Minimizing Δ by varying A and B leads to:

$$\underline{\underline{H}}\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{1}{\sigma_i^2} x_i \\ \sum \frac{1}{\sigma_i^2} x_i & \sum \frac{1}{\sigma_i^2} x_i^2 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sum \frac{1}{\sigma_i^2} y_i \\ \sum \frac{1}{\sigma_i^2} x_i y_i \end{bmatrix}$$

In this case, the inverse matrix, H^{-1}, is:

$$\frac{1}{\sum \frac{1}{\sigma_i^2} \sum \frac{1}{\sigma_i^2} x_i^2 - \left(\sum \frac{1}{\sigma_i^2} x_i\right)^2} \begin{bmatrix} \sum \frac{1}{\sigma_i^2} x_i^2 & -\sum \frac{1}{\sigma_i^2} x_i \\ -\sum \frac{1}{\sigma_i^2} x_i & \sum \frac{1}{\sigma_i^2} \end{bmatrix}$$

Restricting ourselves to the example shown in Figure 7, we find that $H_{22}^{-1} = \dfrac{4\sigma^2}{n\left(x_b - x_a\right)^2}$.

More generally,

$$\sigma_A^2 = H_{11}^{-1}$$
$$\sigma_B^2 = H_{22}^{-1}$$

and

$$\sigma_{q(A,B)}^2 = \left(\frac{\partial q}{\partial A}\right)^2 H_{11}^{-1} + \frac{\partial q}{\partial A}\frac{\partial q}{\partial B} H_{12}^{-1} + \frac{\partial q}{\partial B}\frac{\partial q}{\partial A} H_{21}^{-1} + \left(\frac{\partial q}{\partial B}\right)^2 H_{22}^{-1}$$

If the data must be modeled by the linear sum of several arbitrary foundations:

$$y_i = a_1 f_1\left(x_i\right) + a_2 f_2\left(x_i\right) + \cdots + a_n f_n\left(x_i\right)$$

then the coefficients, $\{a_i\}$, can be determined by solving the matrix equation:

$$\begin{bmatrix} \sum w_i f_1^2\left(x_i\right) & \sum w_i f_1\left(x_i\right) f_2\left(x_i\right) & \cdots \sum w_i f_1\left(x_i\right) f_n\left(x_i\right) \\ \sum w_i f_2\left(x_i\right) f_1\left(x_i\right) & \sum w_i f_2^2\left(x_i\right) & \cdots \sum w_i f_2\left(x_i\right) f_n\left(x_i\right) \\ \vdots & \vdots & \vdots \\ \sum w_i f_n\left(x_i\right) f_1\left(x_i\right) & \sum w_i f_n\left(x_i\right) f_2\left(x_i\right) & \cdots \sum w_i f_n^2\left(x_i\right) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum w_i f_1\left(x_i\right) y_i \\ \sum w_i f_2\left(x_i\right) y_i \\ \vdots \\ \sum w_i f_n\left(x_i\right) y_i \end{bmatrix}$$

The variances for each coefficient, a_j, is given by $\sigma_{a_j}^2 = H_{jj}^{-1}$ and the variance for functions of $\{a_j\}$ is given by:

$$\sigma_{q(a_1, a_2, \dots, a_n)}^2 = \sum_{i,j} \frac{\partial q}{\partial a_i}\frac{\partial q}{\partial a_j} H_{ij}^{-1}$$

Note that the regression matrix defined above is, by construction, symmetric. It can also be shown to be positive-definite. These criteria are sufficient to satisfy the conditions for using the Cholesky decomposition for finding the inverse matrix (see "Numerical Recipes"). Numerical computation issues are often critical if the set of model functions, $\{f_i(x)\}$, is large or the functions are similar in shape. Use of double precision floating point representations is essential. It may also be helpful to select model functions that are orthogonal in the sense that $\int f_i(x) f_j(x) dx = \delta_{ij}$.

Occasionally, you may also need to include constraints on the fitted coefficients, for example $\sum a_i = 1$. These can be incorporated via the method of LaGrangian multipliers as additional linear equations.

The expanded regression matrix no longer satisfies the requirements for the Cholesky decomposition and you must turn to the Singular Value Decomposition (SVD) instead (see "Numerical Recipes").

As long as the model function can be represented by linear sums of well-defined functions, unique solutions can be obtained by matrix inversion with estimates for parameter uncertainty extracted from the appropriate elements of the inverse of the regression matrix. Unfortunately, a large class of problems exist in which the basic model functions depend on parameters in highly non-linear ways. An example is:

$$y_i = Ae^{-\frac{(x_i - B)^2}{2C^2}}$$

The *B* and *C* parameters cannot be computed by any single pass algorithm. For such problems, one must provide initial guesses followed by a search based on trial-and-error that minimizes the square error. In Excel, this is generally accomplished with the Solver add-in. On other platforms, algorithms such as the downhill simplex method (AMOEBA in the Numerical Recipes library) are robust procedures that will usually find the desired optimal values. However, one should always be aware that unlike the situation with linear parameters, the sum of square errors can exhibit local minima in parameter space which are not necessarily the global minima. Thus, the numerical search may find a solution but not the best possible solution.

Estimating the error of non-linear parameters is also much messier. In general, the most reliable technique is performed by simulation of the original data set followed by the identical procedures for estimation of the non-linear parameters.

Problem 7: In Physics 141, an experiment is performed with a dropping ball to determine the local gravitational acceleration, *g*. Assume the ball is dropped at $t = 0$ and the y-coordinate is measured in the downward direction. The standard deviation for the vertical measurements is uniformly 0.012 meters (12 mm). From the following measurements, determine *g* (vertical measurements in meters).

t	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
y	0.463	0.483	0.578	0.652	0.747	0.879	1.051	1.224	1.422	1.671

Also determine the probable error in *g* from the information provided.

Fitting smooth curves to data often requires a certain amount of judgment, particularly if there is no strong physical reason to choose one set of functions over another. One choice when all that is required is smoothness is the cubic spline, an example of which is shown in Figure 2 (the orange curve).

Tests for goodness of fit:

If the data is well described by the chosen model function, one would expect that the average deviation between each data point and the fitted curve would be approximately one standard deviation. If there are *n* data points, this would lead to the expectation that the total square error:

$$\sum \frac{1}{\sigma_i^2}\left(y_i - f\left(x_i\right)\right)^2$$

would be approximately *n* where $f\left(x_i\right) = a_1 f_1\left(x_i\right) + a_2 f_2 + \cdots + a_k f_k\left(x_i\right)$.

In fact, we know that if the number of fitting parameters, *k*, is exactly equal to the number of data points, *n*, the fitted curve will go exactly through each observation and the sum of square errors will be zero. Thus, what really matters is the number of degrees of freedom defined by

$$\nu = n - k$$

and we expect that the total square error is, on average, ν instead of *n*.

The discussion above suggests how we can assess whether a fit to data is good, bad, or indifferent. If we have ν independent, normally distributed quantities, $\{x_i\}$, each with unit variance and zero mean, the sum, $Q = \sum_{i=1}^{\nu} x_i^2$, is distributed according to the χ_ν^2 distribution. The probability peaks at $Q = \nu - 2$ for $\nu \geq 2$ and approaches a Gaussian as ν becomes large (another example of the Central Limit Theorem).

A graph of the differential probability distribution function for χ_ν^2 is shown below in Figure 8. Note the similarity of the curves to the Poisson distribution shown in Figure 6. The functional forms are closely related.

The ratio of square error to number of degrees of freedom is often quoted as a measure of goodness of fit. If this is markedly different from unity, your data may have significant problems or your choice of fitting curve was inappropriate. Significance levels for the χ^2 statistic are usually available on most computing platforms.

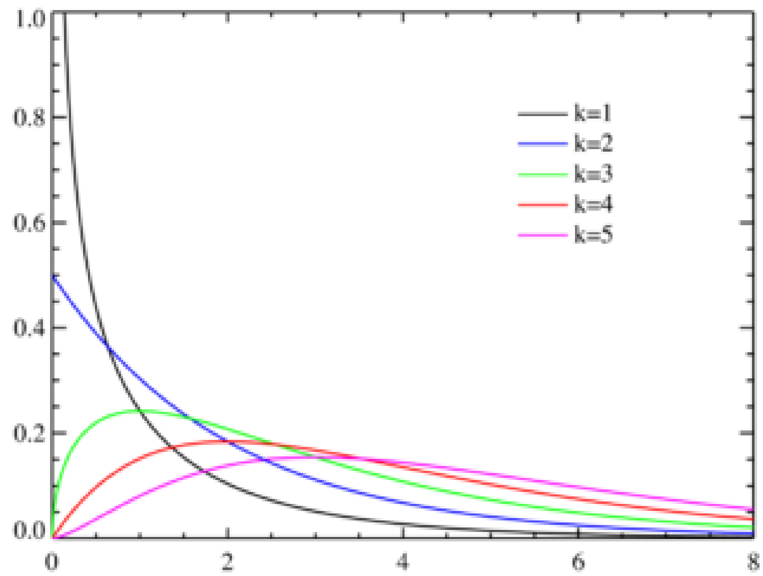

Figure 8. The differential probability function for χ_k^2 for $k=1$ to $k=5$. (Graph obtained from http://en.wikipedia.org/wk/Chi-squared_distribution.)

Probem 8: Compute the χ^2 per degree of freedom for the fitted curve of $y(t)$ for Problem 7.

Testing the similarity of statistical distributions:

In the previous lecture, I presented some methods for evaluating whether two values were likely to be the same or different. That problem can be generalized to the question of whether two probability distribution functions are similar or not. These problems are not likely to arise in Physics 441/442 but it is useful to know that their are well-defined procedures to make such tests. The basis for such comparisons is the cumulative probability distribution function defined by:

$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

Suppose that we wish to compare two such cumulative distributions, $F_1(x)$ and $F_2(x)$. The Kolmogorov-Smirnov test uses the maximum of $|F_1(x) - F_2(x)|$ as a statistic while the Smirnov-Cramér-von Mises test operates on the integral of the squared difference,

$$W^2 = \int_{-\infty}^{\infty} \left(F_1(x) - F_2(x)\right)^2 f(x)\,dx$$

Further information can be found in "Statistical Methods in Experimental Physics", 2nd Edition, Frederick James and "Probability and Statics in Experimental Physics", Byron P. Roe.

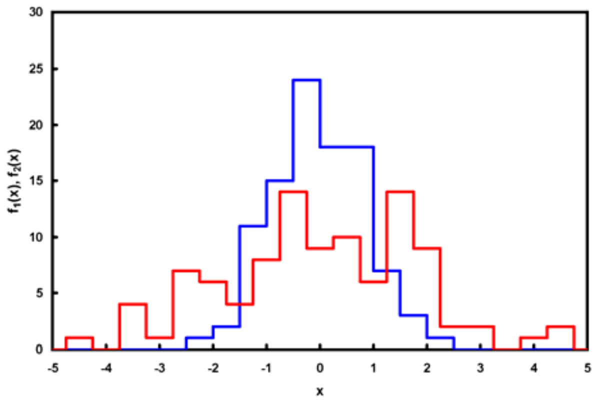

Figure 9. Histogram of two distributions taken from a Gaussian population with σ = 1.0 (blue) and σ = 1.8 (red).

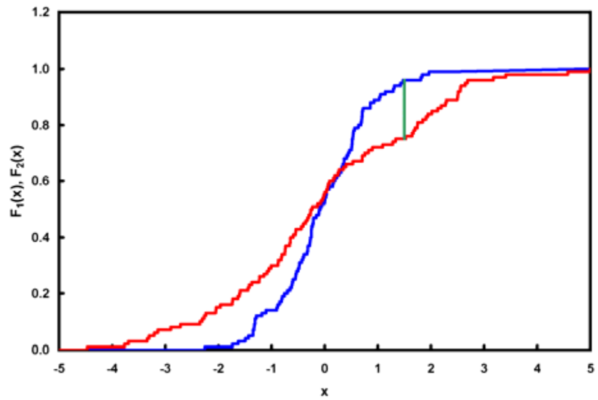

Figure 10. The cumulative distributions for the two samples shown in Figure 9. The vertical green line indicates the maximum separation between $F_1(x)$ and $F_2(x)$.

The graphs shown in Figures 9 and 10 above depict the distribution of 100 observations drawn from Gaussian parent populations with mean of zero and two different values of the variance. The right-hand plot illustrates the application of the Kolmogorov-Smirnov test. In this case, the probability of seeing such a large deviation between the two cumulative distributions is approximately 2.4%.

Signal detection:

Although outside the scope of the experiments performed in Physics 441/442, there are a variety of techniques used in signal analysis and detection that rely on statistical methods similar to those described previously. One of the most interesting of these is the *matched filter*. In this case, it is assumed that one is looking for a time-dependent signal with a well-defined functional form. It is a fairly straight-forward task to devise an optimal weighting scheme that will maximize the signal-to-noise ratio, an important problem for sonar or radar. These techniques easily map to similar problems for two-dimensional image detection and recognition. Finally, the Fast Fourier Transform (FFT) should be mentioned because of its importance for detecting harmonic signals in the face of a variety of noisy backgrounds.

Notes on Statistics

Lecture IV

Carl W. Akerlof

The abuse of the coefficient of distribution, R^2:

A persistent misconception of folks who use Microsoft *Excel* or similar applications is the significance of the coefficient of distribution, R^2, or its square root, R, the coefficient of correlation. These two statistics are appropriate for rat psychologists who wish to demonstrate that there is some connection between two observables although the underlying source of correlation is ill-defined or unknowable. These quantities are incapable of answering the question of whether numerical data are consistent with a particular theoretical prediction. To understand this problem, we start with the definition of R^2:

$$R^2 = 1 - \frac{\sum (y_i - f(x_i))^2}{\sum (y_i - \overline{y})^2}$$

where y_i is the dependent observable associated with the independent variable, x_i, and $f(x_i)$ is the model function that is postulated to describe the correlation of x and y. $\overline{y}$ is the arithmetic mean of the data set, $\{y_i\}$. This definition of R^2 is unambiguous if the functional relation defined by $f(x)$ is linear. For more complex mathematical forms, this convention is not unique. For *Excel* aficionados dealing with power-law, exponential or logarithmic behaviors, the computed value for R^2 is obtained from the linearization instituted by the appropriate transformation of variables. Note that the definition for R^2 is independent of the statistical errors for the dependent variable, y_i. A consequence is shown below for two identical data sets with the errors in Figure 11b five times smaller than in Figure 11a. The fit of the data in Fig. 11a is quite good – the χ^2 statistic shows a high probability that the data is adequately described by a straight line. For Fig. 11b, the situation is just the opposite – the χ^2 statistic indicates a vanishingly small probability that the data is described by such a linear relationship. The R^2 value for

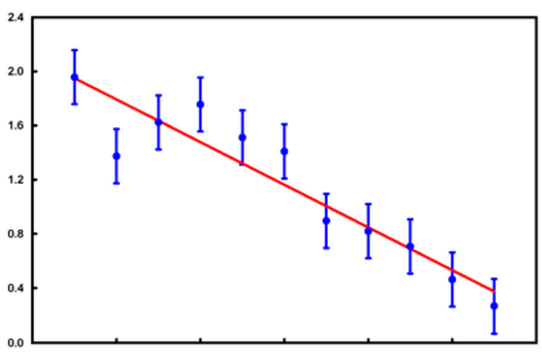

Figure 11a. The straight line fits the data with a χ^2 of 9.485 for 9 degrees of freedom. The probability of this value or larger for χ^2 occurring by chance alone is 0.39. The coefficient of distribution, R^2, is 0.8778.

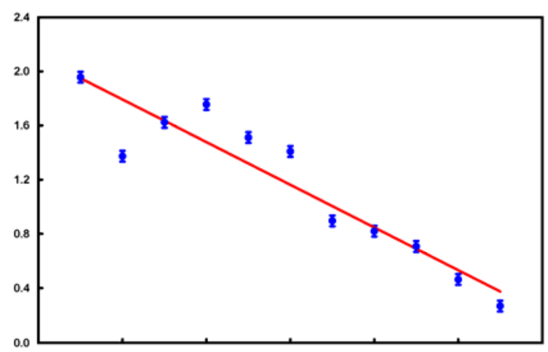

Figure 11b. The straight line fits the data with a χ^2 of 237.126 for 9 degrees of freedom. The probability of this value or larger for χ^2 occurring by chance alone is 5.2×10^{-46}. The coefficient of distribution, R^2, is 0.8778.

these two cases is, of course, identical. This example corresponds to the usual situation in physics – one is trying to prove the validity of a well-defined mathematical relationship that is expected to describe all the relevant data, not just some special subset. For the data shown in Fig. 11b, there must be something significant going on that makes the data depart so clearly from the linear hypothesis. This is the engine by which new discoveries are made.

This begs the question of why anyone would be interested in the coefficient of distribution if it cannot validate a theoretical hypothesis. For many pursuits, particularly in the social sciences, the patterns of behavior are complex and multi-variate. Values of R or R^2 are useful probes of whether there are any correlations that bear more careful investigation and thus these statistics are valuable as tools for data exploration. They do not, by themselves, prove that the correlations are more than statistical flukes. Thus, such techniques are important when the understanding of a phenomenon is exceedingly flimsy. Such occasions can sometimes arise in the physical sciences but are not representative of the problems in Physics 441/442. One should also be a bit skeptical of the significance of the numerical value of R^2. As a numerical experiment, I computed R^2 for linear fits to ensembles of three uniformly spaced data points with identical Gaussian distributions. The differential probability distribution function is shown in Figure 12 below. Note that the PDF tends to peak for both endpoints of R^2.

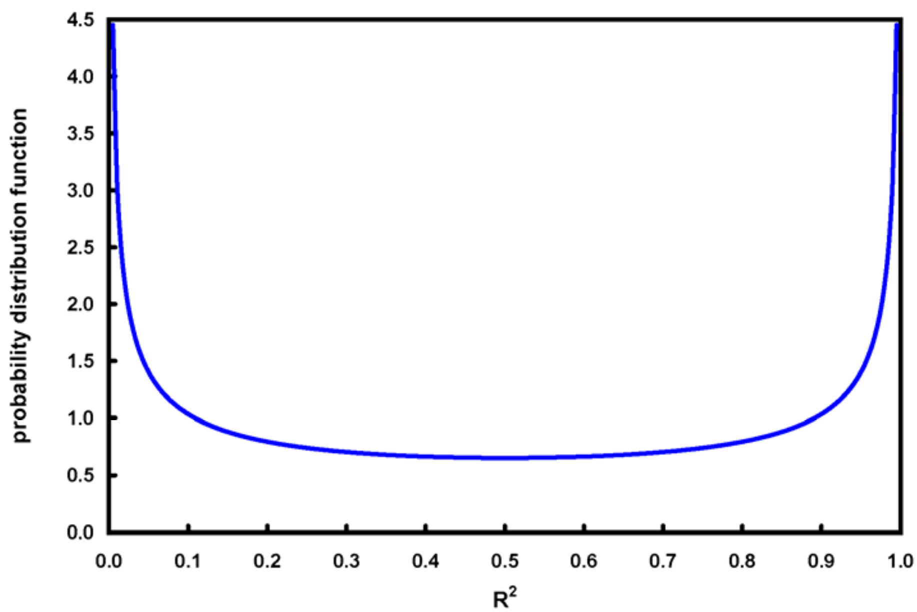

Figure 12. Differential probability distribution function for R^2 for three uniformly spaced Gaussian distributed data points.

The relation between χ^2 and parameter variance:

As previously discussed, one of the most common statistical operations on data is extraction of the parameters of a model function, usually by the method of least squares. If the mathematical model is a linear function of the free parameters, the variance (and co-variance) of the parameters are elements of the inverse of the regression matrix. For more complex mathematical forms, other means are required to determine the accuracy of the inferred parameters. The most obvious procedure is the *Monte Carlo method*. The data are simulated by a distribution function over a set of random numbers that accurately mimics the behavior of the experimental measurements and the variances of the extracted parameters are obtained after performing hundreds or thousands of numerical reproductions of the actual data set. This is conceptually simple but can often require a fair amount of programming. For a complex problem where the behavior of the least square fits is not well understood, it is usually the most reliable procedure.

For a number of problems, a simpler method to estimate the parameter variance is to explore the behavior of the χ^2 statistic as a function of the parameter value. In the following example, we will look at the variance of the mean of a sample, something that we already know how to compute directly:

$$\bar{x} = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

$$\sigma_{\bar{x}}^2 = \frac{\sum \frac{\sigma_i^2}{\sigma_i^4}}{\left(\sum \frac{1}{\sigma_i^2}\right)^2} = \frac{1}{\sum \frac{1}{\sigma_i^2}}$$

If the data is simply the sum of the mean, $\bar{x}$, and a random error, the χ^2 statistic is:

$$\chi^2 = \sum \frac{\left(x_i - \bar{x}\right)}{\sigma_i^2}$$

If the mean is replaced by a slightly different value, $\bar{x}' = \bar{x} + \delta$, the χ^2 will increase by some amount that we will denote by *n*. Thus:

$$\chi^2 + n = \sum \frac{\left(x_i - \bar{x}\right)^2}{\sigma_i^2} + n = \sum \frac{\left(x_i - \bar{x}'\right)^2}{\sigma_i^2} = \sum \frac{\left(x_i - \left(\bar{x} + \delta\right)\right)^2}{\sigma_i^2}$$

To find the relation between *n* and δ, cancel identical terms on either side of the equation:

$$n = -2\delta \sum \frac{x_i}{\sigma_i^2} + 2\delta\, \bar{x} \sum \frac{1}{\sigma_i^2} + \delta^2 \sum \frac{1}{\sigma_i^2}$$

The first two terms on the right-hand side also cancel so that:

$$\delta^2 = \frac{n}{\sum \frac{1}{\sigma_i^2}} = n\,\sigma_{\bar{x}}^2$$

Thus, if you change $\bar{x}$ by one standard deviation, the value increases by one; if $\bar{x}$ changes by $2\sigma_{\bar{x}}$, χ^2 increases by four, etc. This can be generalized to more complex data models. To find the variance of a particular parameter, change its value to see how it affects χ^2. A word of warning: for fits with non-linear parameter behavior, the χ^2 may have complex minima. Thus, it is best to explore what happens if $n = 1, 4, 9, \ldots$ to however far it is important for your conclusions.

Appendix B

Probability Distribution Functions:

$$\frac{dp}{dx} = f(x) \; ; \; \int_{x_{\min}}^{x_{\max}} f(x)dx = 1$$

Cumulative distribution function:

$$F(x) = \int_{x_{\min}}^{x} f(x')dx'$$
$$F\left(x_{\min}\right) = 0 \; ; F\left(x_{\max}\right) = 1$$

Population statistics:

Mean: $\left\langle x \right\rangle = \int_{x_{\min}}^{x_{\max}} x\, f(x)\; dx$

Variance: $\sigma_{\mathrm{x}}^2 = \left\langle x^2 \right\rangle - \left\langle x \right\rangle^2 = \int_{x_{\min}}^{x_{\max}} x^2\, f(x)dx - \left\langle x \right\rangle^2$

Standard Deviation: $\sigma_x = \sqrt{\sigma_x^2}$

Sum of variables: $x_t = \sum_{i=1}^{n} x_i$

Mean: $\left\langle x_t \right\rangle = \sum_{i=1}^{n} \left\langle x_i \right\rangle$

Variance: $\sigma_{x_t}^2 = \sum_{i=1}^{n} \sigma_{x_i}^2$

Sample Statistics:

Sample Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample Variance: $\sigma_s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$

Sample Standard Deviation: $\sigma_s = \sqrt{\sigma_s^2}$

Mean of Sample Means: $\left\langle \overline{x} \right\rangle = \left\langle x \right\rangle$

Mean of Sample Variances: $\left\langle \sigma_s^2 \right\rangle = \sigma_x^2$

Variance of Sample Means: $\sigma_{\overline{x}}^2 = \frac{1}{n}\sigma_x^2$

Variance of Sample Variances:

$$\sigma^2_{\sigma_s^2} = \frac{1}{n}\langle x^4\rangle - \frac{4}{n}\langle x^3\rangle\langle x\rangle - \frac{n-3}{n(n-1)}\langle x^2\rangle^2$$
$$+\frac{4(2n-3)}{n(n-1)}\langle x^2\rangle\langle x\rangle^2 - \frac{2(2n-3)}{n(n-1)}\langle x\rangle^4$$

For Gaussian distributions; $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$

Variance of sample variances: $\sigma^2_{\sigma_s^2} = \frac{2\sigma^4}{n-1}$

Standard deviation of sample variance: $\sigma_{\sigma_s^2} = \sqrt{\frac{2}{n-1}}\ \sigma^2$

Variance of sample standard deviations: $\sigma^2_{\sigma_s} = \frac{\sigma^2}{2(n-1)}$

Standard deviation of sample standard deviations: $\sigma_{\sigma_s} = \frac{\sigma}{\sqrt{2(n-1)}}$

Variance of median: $\sigma^2_{x_{1/2}} = \frac{1}{4nf^2\left(x_{1/2}\right)}$; $F(x_{1/2}) \equiv \frac{1}{2}$

For Gaussian distribution: $\sigma^2_{x_{1/2}} = \frac{\pi}{2n}\sigma^2$

Robust estimate of dispersion: Inter-Quartile Difference (IQD)

Assume $F(x_{1/4}) \equiv \frac{1}{4}$; $F(x_{3/4}) \equiv \frac{3}{4}$
Then: $\mathrm{IQD} = x_{3/4} - x_{1/4}$

For a Gaussian distribution:

$$\frac{\mathrm{IQD}}{q^*} = \sigma \quad \text{where} \quad q^* = 1.3489798\ ; \quad \frac{1}{2} = \frac{1}{\sqrt{2\pi}}\int_{-q^*/2}^{+q^*/2} e^{-\frac{x^2}{2}}dx$$

$$\sigma^2_{(\mathrm{IQD}/q^*)^2} = \frac{2\pi}{n}\ \frac{e^{\left(\frac{q^*}{2}\right)^2}}{q^{*2}}\sigma^4$$
$$= 2.7209\frac{2\sigma^4}{n}$$