Michael E. Peskin
July 11, 2011

Basic Training in Statistics - 1

To test theories of physics, we make predictions and compare these predictions to experimental measurements. Ideally, these measurements will confirm or refute the theory. However, in practice, measurements are always imperfect and even theoretical predictions are not exact. This means that we cannot confirm or reject with perfect certainty. We can only say that a theory is more or less likely to be correct. *Statistics* is the science of quantifying the uncertainties associated with these statements.

The purpose of these lectures is to take you from 0 to ϵ in statistics. The treatment of statistics in reporting experimental measurements has become increasingly sophisticated, sufficiently so that there is even a language gap between theorists and experimenters. I hope that this introduction will be useful in bridging that gap.

Useful elementary and practical references on statistics and the use of statistics in elementary particle physics include:

- The sections of the *Review of Particle Properties* on Probability and Statistics.
- G. Cowan, *Statistical Data Analysis.* (Oxford, 1998)
- L. Lyons, *Statistics for Nuclear and Particle Physics.* (Cambridge, 1986)
- W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes* (Cambridge, 2007), esp. Chapter 14.

Probability

To discuss statistics, we need some basic notions of probability. I will not explain probability philosophically, but it is useful to enuciate the basic properties of probabilities. To an event or outcome a, we associated a probability $p(a)$, a real number in the range

$$0 \leq p(a) \leq 1$$

If a and b are independently generated events,

$$p(a \oplus b) = p(a)\, p(b)$$

If $\{a_i\}$ is a complete set of mutually exclusive events,

$$\sum_i p(a_i) = 1$$

If a and b are not independent, we can define the *conditional probability* of a given b. This is the probability that, if b is known to occur, a also occurs. This probability is given by

$$p(a|b) = p(a \oplus b) / p(b)$$

If $\{b_i\}$ is a complete set of mutually exclusive events,

$$\sum_i p(a|b_i)\, p(b_i) = \sum_i p(a \oplus b_i) = p(a)$$

Conditional probabilities will play a central role in this lecture.

From the expression above,

$$p(a|b)\, p(b) = p(b|a)\, p(a)$$

This can be rewritten as *Bayes' Theorem*:

$$p(a|b) = \frac{p(b|a)\, p(a)}{p(b)}$$

This is an important identity, with a significance that I will discuss in detail below.

Often, outcomes a are parametrized by a continuous variable x. Then the probability for the outcome to be a value between x and $x + dx$ is expressed as

$$p(x)\, dx$$

normalized to

$$\int_{-\infty}^{\infty} dx\; p(x) = 1$$

The function $p(x)$ is called the *probability distribution function* or *pdf*. The integral of this function is the *cumulative distribution*. It is useful to define *lower* and *upper* cumulative distributions $P(x)$ and $Q(x)$

$$P(x) = \int_{-\infty}^{x} dx\; p(x) \qquad Q(x) = \int_{x}^{\infty} dx\; p(x)$$

The function $P(x)$ is the probability that the outcome y is less than x. P and Q obey

$$\lim_{x\to\infty} P(x) = 1 \qquad \lim_{x\to\infty} Q(x) = 0 \qquad Q(x) = 1 - P(x)$$

The *expectation value* of a function of x is

$$\langle f(x) \rangle = \int dx\, f(x)\, p(x)$$

By definition, $\langle 1 \rangle = 1$. Some important expectation values are the *mean*

$$\bar{x} = \langle x \rangle = \int dx\, x\, p(x)$$

and the *variance*

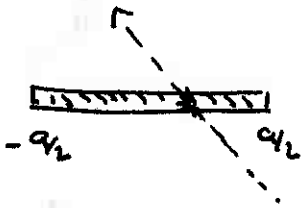
$$\mathrm{Var}(x) = \langle (x - \bar{x})^2 \rangle = \langle x^2 \rangle - \bar{x}^2$$

Expectation values and cumulative distributions are defined also, in the obvious way, for processes with discrete outcomes.

It is worth recalling some distributions that arise often in physics:

Box distribution:

$$p(x) = \begin{cases} \frac{1}{a} & -\frac{a}{2} < x < \frac{a}{2} \\ 0 & \text{otherwise} \end{cases}$$



This describes the location of a hit in a sensor of size a.

$$\langle x \rangle = 0 \qquad \mathrm{Var}(x) = \frac{a^2}{12} = (0.29a)^2$$

Binomial distribution:

$$p(k) = \binom{n}{k} \wp^k (1-\wp)^{n-k}$$

This describes the probability of k positive results in n trials for a process with probability $\wp$ of obtaining a positive result in each trial.

$$\langle k \rangle = n\wp \qquad \mathrm{Var}(k) = n\wp(1-\wp)$$

Poisson distribution:

$$p(k) = \frac{1}{k!} \lambda^k e^{-\lambda}$$

This describes the probability of finding k events from a process that generates events independently and a function of time.

$$\langle k \rangle = \lambda \qquad \mathrm{Var}(k) = \lambda$$

Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

This is the canonical statistical "bell-shaped curve".

$$\langle x \rangle = \mu \qquad \mathrm{Var}(x) = \sigma^2$$

We will encounter several more useful pdf's in the course of these lectures. I will tabulate the properties of all of these in Appendix A of these notes.

The Gaussian distribution has a privileged place in statistics. First of all, it is simple to manipulate. Many of the statistical constructions that are difficult to evaluate for a general pdf become trivial to evaluate when the pdf is a Gausssian. The results for Gaussian pdf's are often quoted as *statistical rules* which are sometimes applied blindly to data following more general distributions.

Second, the *Central Limit Theorem* asserts that that pdf for a parameter derived from a number of measurements converges to a Gaussian in the limit in which this number becomes large. Through this logic, Gaussian distributions often arise naturally in the analysis of experimental data.

Here is a simple argument for the Central Limit Theorem: Let x_i, $i = 1, \ldots, N$, be random variables with the pdf's $p_i(x_i)$. I will assume here that each of these pdf's has a finite expectation value for any moment of x_i, that is that $\langle x_i^n \rangle$ is bounded. In a careful proof of the Central Limit Theorem, it is only necessary to assume that the mean and variance of x_i are bounded. Shift each x_i to a variable whose mean is zero:

$$y_i = x_i - \langle x_i \rangle$$

I will compute the probability distribution of

$$Y = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \bar{x}_i$$

and show that is converges to a Gaussian as N becomes large.

The pdf for Y is

$$p(Y) = \int dy_1 \cdots dy_N \prod_i p_i(y_i) \; \delta(Y - \sum_i^N y_i)$$

This is a convolution of the distributions $p_i(y_i)$. Represent each of these pdf's by its Fourier transform,

$$p_i(y) = \int \frac{dk}{2\pi} e^{iky} \tilde{p}_i(k)$$

Then

$$\tilde{p}_i(k=0) = \langle 1 \rangle = 1 \qquad -i \frac{d}{dk} \tilde{p}_i(k)\Big|_0 = \langle y_i \rangle = 0$$

$$-\frac{d^2}{dk^2} \tilde{p}_i(k)\Big|_0 = \langle y_i^2 \rangle \quad \text{etc.}$$

We can write $\tilde{p}_i(k)$ in the form

$$\tilde{p}_i(k) = \exp\left[a_i + i b_i k - \tfrac{1}{2} c_i k^2 + \cdots \right]$$

where

$$a_i = 0 \qquad b_i = 0 \qquad c_i = \langle y_i^2 \rangle$$

The coefficients c_i, d_i, ... are all numbers of order 1. Then the Fourier transform of the pdf of Y is

$$\tilde{p}(k) = \prod_i \tilde{p}_i(k) = \exp\left[-\tfrac{1}{2} C k^2 - i D k^3 + \cdots \right]$$

where

$$C = \sum_i c_i \qquad D = \sum_i d_i \qquad \text{etc.}$$

The parameters C, D, ... are of order N. Since C is of order N, $\tilde{p}(k)$ becomes very small already when k is $\mathcal{O}(1/\sqrt{N})$. At this point, $Dk^3 \sim \mathcal{O}(1/\sqrt{N})$, $Ek^4 \sim \mathcal{O}(1/N)$, and so these and all successive terms are increasingly irrelevant. Thus, as $N \to infty$, $\tilde{p}(k)$ is increasingly well approximated by

$$\tilde{P}(k) = \exp\left[-\tfrac{1}{2}Ck^2\right]$$

That is

$$P(Y) = \frac{1}{\sqrt{2\pi C}} \exp\left[-\frac{Y^2}{2C}\right]$$

where

$$C = \sum_i \mathrm{Var}(y_i)$$

I gave this argument explicitly to show its great generality. However, it contains a problem, which is that the convergence to a Gaussian is slowest in the tails of the distribution. We will see shortly that statistical inference is crucially tied to the behavior of tails of distributions. So it is important in each case to think carefully about the real expectation for the tail of the relevant distribution before claiming that a hypothesis is excluded with high probability.

For problems like this, and to check other inferences from formal statistics, it is useful to create a set of simulated experiments (called *pseudo-experiments*). Computer time is cheap, and experimentation is valuable. You might be surprised by the results.

Hypothesis testing

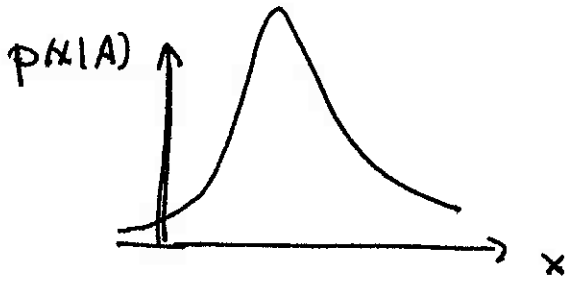
With these preliminaries, we are ready to begin discussing statistics proper. The most elementary type of statistical inference is a *hypothesis test*: Given model A and datum x, is A compatible with x? Here x may be a single measurement or a set of measurements, and A might also be parametrized by some discrete or continuous parameters.

The basic object that we use to make such inferences is the *likelihood function*

$$p(x|A)$$

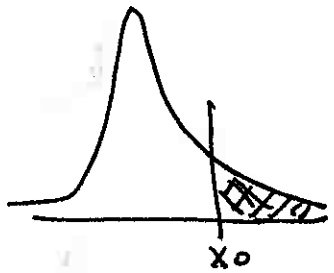
The function $p(x|A)$ is the probability, given that model A is correct, to find the datum x. Typically, it is conceptually straightforward to compute the likelihood function. What is not so straightforward to using the likelihood function to give back information about A. I will denote the value of x actually obtained in an experiment as x_0. It is clear that a small value of $p(x_0|A)$ implies that A is unlikely. But, how do we quantify this?

For simplicity, take x to be one number. A plot of $p(x|A)$ will have the form



If x_0 is a value on the tail of $p(x|A)$, define

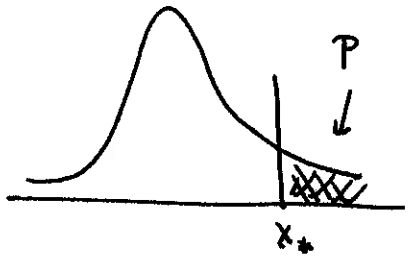
$$\alpha = \int_{x_0}^{\infty} dx\, p(x|A) = Q(x_0|A)$$



where Q is the upper cumulative distribution of $p(x|A)$. If the likelihood function has been computed correctly and the hypothesis A were correct, then if the experiment were repeated many times, α would be the probability for the outcome to be equally or less compatible with A than x_0. If α is small, say, 1%, it is unlikely that the incompatibility of x_0 with the center of the distribution $p(x|A)$ is just a matter of chance. In this case, we say that the hypothesis A is "excluded at the 1% confidence level".

More generally, we can define a criterion of incompatibility beyond which we would reject the hypothesis A, defined by some boundary x_*. The value

$$p = Q(x_* | A)$$



is called the *p-value* of the test. Alternatively, we speak of a test with a *confidence level* given by $(1 - p)$. A p-value of 10% corresponds to a 90% confidence level test.

It is instructive to write out the p-values for a one-dimensional Gaussian distribution,

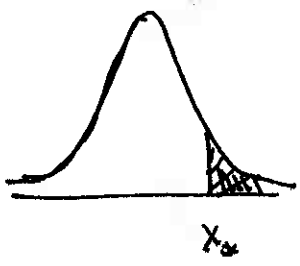
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Here

$$Q(x_*) = \int_{x_*-\mu}^{\infty} dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{2} \operatorname{erfc}\left(\frac{x_*-\mu}{\sqrt{2}\sigma}\right)$$

Then

x_*	p-value
1σ	15.9%
2σ	2.3%
3σ	0.14%
5σ	2.9×10^{-7}



The p-value decreases very rapidly with the deviation from the center of the Gaussian. A p-value of 5% (95% confidence exclusion) corresponds to a deviation of only 1.64σ from the maximum value.

p-value	x_*
5%	1.64σ
1%	2.33σ
10^{-3}	3.09σ

Of course, this depends on the distribution having an accurate Gaussian fall-off in the tails. Note also that the easiest way to obtain a very small p-value is to underestimate the value of σ.

A Gaussian distribution in N dimensions has the general form

$$\left[\frac{1}{\det 2\pi V}\right]^{\frac{1}{2}} \exp\left[-\frac{1}{2}\sum_{ij}(x_i-\mu_i)(V^{-1})_{ij}(x_j-\mu_j)\right]$$

If we shift each variables x_i to its mean μ_i, $y_i = x_i - \mu_i$, and then diagonalize V,

$$V = U\begin{pmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2\end{pmatrix}U^{-1}$$

rotate coordinates, and rescale the y_i to remove the σ_i^2, this distribution takes the form

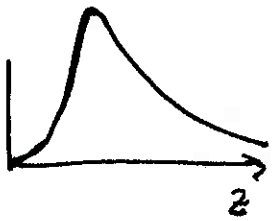
$$p(w) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_i w_i^2}$$

Now go to spherical coordinates in the w_i. The variable

$$z = \sum_i^1 w_i^2$$

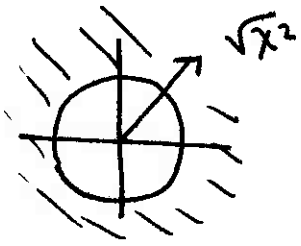
is distributed according to the χ^2 *distribution with N degrees of freedom*

$$p(z) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$



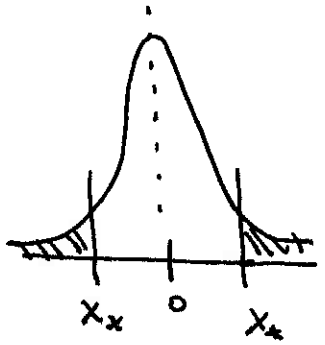
The p-values for a χ^2 distribution with N d.o.f. are

p-value	N=1	2	3	$\chi^2 / \sqrt{\chi^2}$ (6)
31.7%	1/1	2.3/1.5	3.5/1.9	
10%	2.7/1.6	4.6/2.1	6.3/2.5	
5%	3.8/1.9	6.0/2.4	8.0/2.8	
1%	6.6/2.6	9.2/3.0	11.3/3.4	



The case $N = 1$ corresponds to a 2-sided test with a Gaussian distribution, for which the p-values are double those given above:

x_+	p-value
1σ	31.7%
2σ	4.6%
3σ	0.27%



Let's return to our general considerations on testing a hypothesis A with a datum x. An interesting philosophical question – which I have not yet addressed – is: What is the probability of the hypothesis A given a measurement x_0? The problem here is to invert the implication given in the conditional probability $p(x_0|A)$. There are two different canonical answers to this question. The difference exposes a basic controversy in the interpretation of measurements.

Frequentists say that it is not possible conceptually to assign a probability to A. Instead, we use the p-value to quantify thye compatibility of the data with A. If the p-value is small, A is likely to be incompatible with x_0. However, the p-value should not be interpreted as the probability of A being true.

Bayesians say that it is meaningful to assign a probability to A. And, they have a method to do this in practice, by inverting the conditional probability in the likelihood using Bayes' theorem:

$$p(A|x_0) = \frac{p(x_0|A)\, p(A)}{p(x_0)}$$

To evaluate this formula, we take $p(x_0) = 1$, since x_0 was the actual outcome of the experiment. Then the probability of A given x_0 is proportional to the likelihood

$p(x_0|A)$. But, there is another ingredient, the quantity $p(A)$ which gives the *a priori* probability, before the measurement was made, that A was correct. We call $p(A)$ the *prior probability* or, simply, the *prior*, and $p(A|x_0)$ the *posterior probability.* The Bayesian interpretation of the equation is that the measurement of x transforms our understanding of the probability of A before the measurement by adding the information given in the measured value. A difficulty with this interpretation, which goes along necessarily with the Bayesian viewpoint, is that the prior $p(A)$ is subjective, reflecting in some way our incomplete state of knowledge.

The simplest way to apply the Bayesian formula occurs in the case where we have a family of hypotheses $A(a)$ depending on a continuous parameter a. For example, x might be the peak of a reconstructed mass distribution, and a might be the particle mass in the theory that describes this peak. Before the measurement, we might say that we are totally ignorant of the value of a, and that therefore $p(a)$ is a flat distribution. After the measurement, the probability of a is proportional to $p(x_0|A(a))$. The correctly normalized posterior probability of a would then be

$$p(a|x_0) = \frac{p(x_0|a)}{\int da'\, p(x_0|a')}$$

Please note that this distribution depends on the choice of the parameter a. If we parametrize the family of theories in terms of another parameter b, we might apply a prior that is flat in b. This gives a different result for the posterior probability distribution:

$$p(a|x_0) = \frac{p(x_0|a)\,\frac{db}{da}(a)}{\int da'\, p(x_0|a')\,\frac{db}{da'}}$$

You see the problem: Frequentists think that Bayesians are excessively subjective; Bayesians think that frequentists have blinders on.

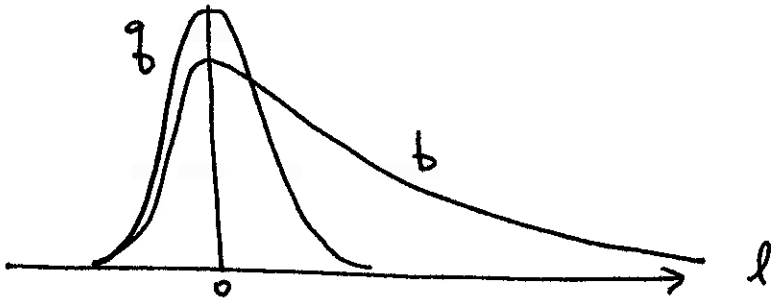
Particle theorists feel that there is real meaning to questions about the probability of hypotheses. For example, they feel that the question "What is the probability that the Higgs boson mass is between 114 GeV and 120 GeV?" can be well posed. Thus, typically, particle theorists are Bayesians. Astronomers are also typically Bayesians. The rude explanation is that there are so many uncertainties in astrophysics that it impossible to reach a useful conclusion within help from prior knowledge. On the other hand, particle experimenters tend to be frequentists. A rude explanation is

that 3,000 highly opinionated collaborators could never agree on the choice of a prior. To understand the literature, then, it is necessary to speak both languages. In the remainder of these lectures, you will find arguments from both viewpoints.

A simple application of hypothesis testing is the discrimination of two alternative hypotheses A and B. For example, consider measurements on jets at a high-energy hadron collider. Let ℓ be the location of the best-reconstructed secondary vertex in the jet. Let $A = q$ be the hypothesis that the jet originated from a light quark or gluon and let $B = b$ be the hypothesis that the jet originated from a b quark. We can make a choice between q and b based on the likelihoods

$$p(\ell|q) \qquad p(\ell|b)$$

The two distributions have the form:



Thus, we can assign a jet to be a b jet if $\ell > \ell_*$ for some distance ℓ_*.

With this method of choosing between the hypotheses, the efficiency for identifying a b jet is

$$\varepsilon_b = \int_{\ell_*}^{\infty} d\ell \; p(\ell|b) = Q(\ell_*|b)$$

where, again, $Q(\ell|b)$ is the upper cumulative distribution. The probability to incorrectly identify a light quark jet as a b jet is

$$\varepsilon_l = \int_{\ell_*}^{\infty} d\ell \; p(\ell|q) = Q(\ell_*|q)$$

This is just the p-value of the test that the jet is not a light quark jet. By adjusting ℓ_*, we can tune our sample of b jets for increased efficiency or increased purity.

Bayesians would go a little farther along this line of reasoning. Let $p(q)$, $p(b)$ be the *a priori* probabilities that a jet in the sample being studied is a light quark jet or a b jet. Then, after the measurement of ℓ, the posterior relative probability that this jet is a b jet is

$$\frac{p(\ell|b)\ p(b)}{p(\ell|q)\ p(q)}$$

In a typical jet sample, we would have $p(q) \gg p(b)$, and therefore it would be sensible to choose a large values of ℓ_* to bring the posterior probability closer to 1:1.

Parameter estimation: maximum likelihood

Next, consider the situation in which the value of a measurement x depends on some underlying parameter θ. If we make a series of estimates of x, this should allow us to estimate the correct underlying value of θ. How do we do this?

There is a very general answer to this question that makes the best use of all available information. This is to use the likelihood function as the estimation tool. That is, construct

$$p(x|\theta)$$

and let the estimate of θ be the value $\hat{\theta}$ that maximizes

$$p(x_0|\theta)$$

using the measured value x_0. This choice of $\hat{\theta}$ is the *maximum likelihood estimator* of θ.

Maximum likelihood is a very convenient prescription that automatically answers many basic questions. For example, if we have multiple independent measurements,

the likelihood function for the set of measurements is simply the product of the likelihood functions for each measurement. To estimate θ, we construct

$$p(\{x_i\}|\theta) = \prod_i p(x_i|\theta)$$

and find the maximum of this function over θ.

Maximum likelihood can be applied to unbinned data, but it also tell us how to estimate θ when the data is binned. From the exact likelihood $p(x|\theta)$, we can construct the binned likelihood

$$p(k|\theta) = \int_{\text{bin } k} dx \; p(x|\theta)$$

and then use this latter function in the product formula for the complete likelihood function of the data set.

Here are two instructive examples of maximum likelihood estimation. First, consider the measurement of a particle lifetime based on a set of measured values of the decay time. For a particle of lifetime τ, the decay times follow an exponential distribution

$$p(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$$

If we have N measured decays times, the likelihood function is

$$p(\{t_i\}|\tau) = \prod_i \left(\frac{1}{\tau} e^{-t_i/\tau} \right)$$

It is easiest to deal with the extended product by taking the logarithm to convert the product to a sum. This gives the log likelihood function $\mathcal{L}$. For this case,

$$\mathcal{L} = \sum_{i=1}^{N} \left(\log \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

The maximization of $\mathcal{L}$ is equivalent to the maximization of the original likelihood. Thus, determine $\hat{\tau}$ from the equation

$$0 = \frac{\partial}{\partial \tau} \mathcal{L} = \sum_{i=1}^{N} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = -\frac{N}{\tau} + \frac{\sum_i t_i}{\tau^2}$$

That is

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} t_i$$

This is the expected result that $\hat{\tau}$ is estimated as the average of the observed decay times.

It is interesting to parametrized the distribution of decay times in a different way and work through the analysis again. Let λ be the particle decay rate; then

$$p(t|\lambda) = \lambda e^{-\lambda t}$$

For the data given, the log likelihood function is

$$\mathcal{L} = \sum_{i=1}^{N} (\log \lambda - \lambda t_i)$$

Maximizing this with respect to λ, we find

$$\frac{N}{\hat{\lambda}} - \sum_i t_i = 0$$

or

$$\hat{\tau} = \left[\frac{1}{N}\sum_i t_i\right]^{-1}$$

It is noteworthy that this result is equivalent to our earlier expression for $\hat{\tau}$ by a change of variables. This illustrates an important property of the maximum likelihood estimator. We find the same answer from this method independently of how the underlying model is parametrized.

There is a sense, however, in which one of the two estimation formulae is better than the other. The expectation value of t in a exponential distribution is

$$\langle t_i \rangle = \tau$$

so the expectation value of the estimate

$$\langle \hat{\tau} \rangle = \left\langle \frac{1}{N}\sum_i t_i \right\rangle = \tau$$

agrees with the underlying value τ for any N. On the other hand, since

$$\left[\sum_i t_i\right]^{-1} = \int_0^\infty d\omega \; e^{-\omega \sum t_i}$$

the expectation value of $\hat{\lambda}$ is

$$\langle \hat{\lambda} \rangle = \left\langle \frac{N}{\sum_i t_i} \right\rangle = N \left\langle \int_0^\infty d\omega \prod_{i=1}^{N} e^{-\omega t_i} \right\rangle$$

$$= N \int_0^\infty d\omega \prod_i \int_0^\infty dt_i \lambda e^{-\lambda t_i} e^{-\omega t_i} = N \int_0^\infty d\omega \left(\frac{\lambda}{\lambda+\omega}\right)^N$$

Thus, the estimator $\hat{\lambda}$ is *biased.*

$$\langle \hat{\lambda} \rangle = \frac{N}{N-1} \lambda = \lambda \left(1 + \frac{1}{N} + \ldots\right)$$

For small values of N, it systematically overestimates the value of λ. An estimator of a parameter, however, we obtain it, can be biased in this way. For the maximum likelihood estimator, the estimate does not depend on the coordinate system, but the bias does depend on what coordinates are used.

More generally, in designing an estimator for a parameter θ, we need to minimize both the variance and the bias. The complete mean square error is a combination of these factors:

$$\langle (\hat{\theta} - \theta_0)^2 \rangle = \langle (\theta - \langle \hat{\theta} \rangle + \langle \hat{\theta} \rangle - \theta_0)^2 \rangle$$

$$= \langle (\theta - \hat{\theta})^2 \rangle + (\langle \hat{\theta} \rangle - \theta_0)^2$$

$$= \mathrm{Var}(\theta) + (\mathrm{Bias}(\hat{\theta}))^2$$

The bias of an estimator is not so easy to work out directly. This is a place where pseudoexperiments can give valuable information.

As a second example, we can analyze the estimate of the mean and variance of a Gaussian distribution from data. I will assume that the underlying distribution is

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

If we sample the distribution N times, we obtain measurements x_i. the log likelihood $\mathcal{L}$ for this set of measurements is

$$\mathcal{L} = \sum_{i=1}^{N} \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

We now must maximize this distribution with respect to the parameters μ and σ^2. Differentiating $\mathcal{L}$ with respect to μ, we find

$$0 = \left. \frac{\partial \mathcal{L}}{\partial \mu} \right|_{\hat{}} = + \frac{1}{\hat{\sigma}^2} \sum_i (x_i - \hat{\mu})$$

so that

$$\hat{\mu} = \frac{1}{N} \sum_i x_i$$

as expected. This is an unbiased estimate. Differentiating $\mathcal{L}$ with respect to σ^2, we find

$$0 = \left. \frac{\partial \mathcal{L}}{\partial \sigma^2} \right|_{\hat{}} = \sum_i \left[-\frac{1}{2} \frac{1}{\hat{\sigma}^2} + \frac{(x_i - \hat{\mu})^2}{2(\hat{\sigma}^2)^2} \right]$$

so we find the estimate

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i (x_i - \hat{\mu})^2$$

This makes sense, but it is actually a biased estimate of σ^2. To see this, note that the values x_i form a vector that fluctuates according to a Gaussian distribution in an N-dimensional space. However, since $\hat{\mu}$ is formed from the x_i, the quantity

$$\sum_i (x_i - \hat{\mu})^2$$

subtracts out the fluctuations of the x_i along the direction $(1, \ldots, 1)$. The remaining components of the vector fluctuate according to a Gaussian distribution in a space of $(N-1)$ dimensions and therefore the variance in the length of the vector is

$$(N-1)\,\sigma^2$$

Thus, the unbiased estimator for σ^2 is

$$s^2 = \frac{1}{N-1} \sum_i (x_i - \hat{\mu})^2 \quad \text{with} \quad \hat{\mu} = \frac{1}{N} \sum_i x_i$$

and the estimate $\hat{\sigma}^2$ above consistently underestimates σ^2 for small values of N.

Parameter estimation: least squares

It is often a good approximation to expand the log likelihood function about its maximum. This is especially true if the log likelihood function is a composite from a large number N of samples. The logic is the same as tha in the proof of the Central Limit Theorem: For N large, the quadratic terms in the exponent of the likelihood function drive the likelihood to zero before the higher-order terms have a chance to become important. Then, for the determination of one variable θ, the log likelihood function can be written as

$$\mathcal{L} = -\frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - g(\theta))^2}{2\sigma^2}$$

From here on, I will assume that the form of the likelihood function, including the value of σ^2, is known. We then determine $\hat{\theta}$ by minimizing the quadratic form

$$\sum_{i=1}^{N} \frac{(x_i - g(\theta))^2}{2\sigma^2}$$

The resulting $\hat{\theta}$ is called the *least squares* estimation of θ.

To analyze this further, expand the function $g(\theta)$ about $\hat{\theta}$,

$$g(\theta) = g(\hat{\theta}) + a(\theta - \hat{\theta}) + \cdots$$

The log likelihood becomes

$$\mathcal{L} = -\frac{1}{2} \sum_i \frac{(x_i - g(\hat{\theta}))^2}{2\sigma^2} - \frac{N}{2} (\theta - \hat{\theta})^2 \frac{a^2}{\sigma^2}$$

The cross term vanishes because $\hat{\theta}$ maximizes $\mathcal{L}$. The first term is in $\mathcal{L}$ is generally nonzero. The values of

$$z = \sum_i \frac{(x_i - g(\hat{\theta}))^2}{\sigma^2}$$

follow a χ^2 distribution with $(N-1)$ degrees of freedom. As I will discuss in the next lecture, this can be used to test the adequacy of the model in which only one parameter θ need be adjusted to describe the data. The second term describes the peaking of the likelihood function about the estimated value $\hat{\theta}$ and allows us to estimate the extent to which the likelihood function constrains θ. The dependence is

$$\text{likelihood} \sim e^{-\frac{1}{2} \frac{(\theta - \hat{\theta})^2}{\Sigma^2}}$$

where

$$\Sigma = \frac{1}{\sqrt{N}} \frac{\sigma}{a}$$

The value Σ gives the standard error on the determined value $\hat{\theta}$. I will discuss the meaning of this error in some detail in the second lecture.

It is interesting to generalize the formulae for the quadratic approximation to the log likelihood to multiple correlated measurements and to multiple parameters. I denote the measurements by x_i, $i = 1, \ldots, N$, and the parameters by θ_a, $a = 1, \ldots, M$. Then, to quadratic order in the x_i,

$$\mathcal{L} = -\frac{1}{2} \sum_{i,j=1}^{N} (x_i - g_i(\Theta))(V^{-1})_{ij}(x_j - g_j(\Theta))$$

The matrix V_{ij} is called the *covariance matrix* of the x_i. We would estimate the parameters θ_a by minimizing the sum of squares with respect to the θ_a.

Let $\hat{\theta}_a$ be the values at the minimum and expand

$$g_i(\Theta) = g_i(\hat{\Theta}) + A_{ia}(\Theta - \hat{\Theta})_a + \cdots$$

The values $\hat{\theta}_a$ are given by the M equations

$$A^T_{ai}(V^{-1})_{ij}(x_j - g_j(\hat{\Theta})) = 0$$

Solving these equations, we absorb M dimensions of the fluctuations of the x_i. The value of the log likelihood at the minimum

$$-2\mathcal{L} = \sum_{ij} (x_i - g_i(\hat{\Theta}))\, V^{-1}_{ij}\, (x_j - g_j(\hat{\Theta}))$$

then follows a χ^2 distribution with $(N - M)$ degrees of freedom.

The dependence of the log likelihood on the θ_a is

$$\mathcal{L} = (\text{above}) - \frac{1}{2}\sum_{ij\,ab} (\theta_a - \hat{\theta}_a) A^T_{ai} V^{-1}_{ij} A_{jb} (\theta_b - \hat{\theta}_b)$$

$$= (\text{above}) - \frac{1}{2}\sum_{ab} (\theta_a - \hat{\theta}_a) U^{-1}_{ab} (\theta_b - \hat{\theta}_b)$$

Then the eigenvalues of the matrix

$$U = (A^T V^{-1} A)^{-1}$$

give the standard errors σ^2 for the components of the estimated $\hat{\theta}$.

While we have these formulae on display, it is useful to discuss one further issue, the question of how to combine sets of measurements governed by Gaussian errors. I will consider only the 1-variable case for simplicity. Consider two independent sets of measurements, giving the likelihood functions for θ

$$P_1(x_{10}|\theta) = A_1 e^{-(\theta - \hat{\theta}_1)/2\Sigma_1^2}$$

$$P_2(x_{20}|\theta) = A_2 e^{-(\theta - \hat{\theta}_2)^2/2\Sigma_2^2}$$

The combined likelihood function is

$$P(x_{10}, x_{20}|\theta) = A_1 A_2 \exp\left[-\frac{(\theta - \hat{\theta}_1)^2}{2\Sigma_1^2} - \frac{(\theta - \hat{\theta}_2)^2}{2\Sigma_2^2}\right]$$

This rearranges into

$$P(x_{10}, x_{20}|\theta) = B \exp\left[-\frac{1}{2}\left(\frac{1}{\Sigma_1^2} + \frac{1}{\Sigma_2^2}\right)\theta^2 + \theta\left(\frac{\hat{\theta}_1}{\Sigma_1^2} + \frac{\hat{\theta}_2}{\Sigma_2^2}\right)\right]$$

$$= C \exp\left[-\frac{1}{2}\frac{(\theta - \tilde{\theta})^2}{\tilde{\Sigma}^2}\right]$$

Thus, the combined estimate of θ is

$$\tilde{\theta} = \frac{\left(\frac{\hat{\theta}_1}{\Sigma_1^2} + \frac{\hat{\theta}_2}{\Sigma_2^2}\right)}{\left(\frac{1}{\Sigma_1^2} + \frac{1}{\Sigma_2^2}\right)}$$

and the combined standard error on this estimate is

$$\tilde{\Sigma}^2 = \left(\frac{1}{\Sigma_1^2} + \frac{1}{\Sigma_2^2}\right)^{-1}$$

If the likelihoods are not Gaussian, a more detailed consideration of the form of the likelihood functions is needed. This brings us into the question of how to quote error bounds for estimates with general forms of the probability function, a subject that I will take up in the second lecture.