LBL-16993 Rev.
c.2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

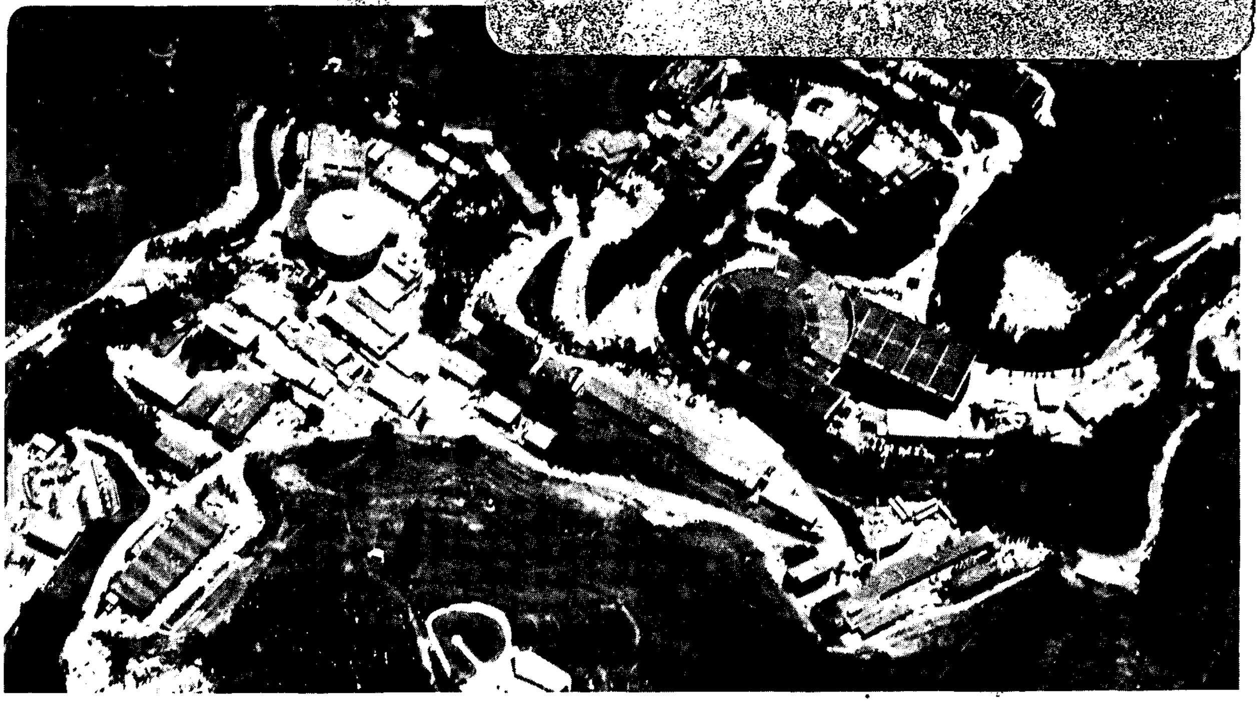
Physics Division

Presented as a course on Statistics,
Imperial College, London, U.K.,
January-April 1983

LECTURES ON PROBABILITY AND STATISTICS

G.P. Yost

June 1985



Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-16993 Rev.
84/HENP/3
June 1985

LECTURES ON PROBABILITY AND STATISTICS

G.P. Yost*

HENP Group, Imperial College, London SW7 2BZ, UK

Work supported in part by the UK Science and Engineering
Research Council and by the US National Science Foundation.
Partial support by the U.S. Department of Energy under
Contract No. DE-AC03-76SF00098

*Permanent address: Physics Department, University of California, Berkeley, CA 94720 USA
and Bldg. 50, Room 308, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720 USA

Table of Contents

INTRODUCTION

These notes are based on a set of statistics lectures delivered at Imperial College to the first-year postgraduate students in High Energy Physics. They are designed for the professional experimental scientist.

We begin with the fundamentals of probability theory, in which one makes statements about the set of possible outcomes of a experiment, based upon a complete *a priori* understanding of the experiment. For example, in a roll of a set of (fair) dice, one understands *a priori* that any given side of each die is equally likely to turn up. From that, we can calculate the probability of any specified outcome.

We finish with the inverse problem, statistics. Here, one begins with a set of actual data (e.g., the outcomes of a number of rolls of the dice), and attempts to make inferences about the state of nature which gave those data (e.g., the likelihood of seeing any given side of any given die turn up). This is a much more difficult problem, of course, and one's solutions often turn out to be unsatisfactory in one respect or another. Hopefully, the reader will come away from these notes with a feel for some of the problems and uncertainties involved. Although there are standard approaches, most of the time there is no cut and dried "best" solution — "best" according to every criterion.

Even though the audience was composed of high energy physicists, the contents should prove to be useful in other fields as well. Since these notes were originally prepared for oral delivery, they may be somewhat terse in some places. The author would appreciate any comments which might result in an improvement. In particular, of course, if any factual errors, misprints, or mistakes in the equations could be quickly caught, it would be a real service.

SECTION A. **PROBABILITY**

Chapter 1. First Principles

The Calculus of Probabilities

Let us perform some experiment which has outcomes. The probability space Ω consists of the set of all possible outcomes or events E. We assign a probability P(E) to each event. We insist on an intuitive meaning for P(E) s.t.(such that), if the identical experiment were to be repeated a large number of times, P(E) would tell us the fraction of times we could expect to find E. This is the frequency definition of probability. Some experiments cannot be repeated, (e.g., a certain patient with a certain disease), but the probability represents in some intuitive sense the chances of each possible outcome.

Axioms of Probability

The following axioms lead to a model for probability that intuition would demand:

1.	$P(\Omega) = 1;$	Something has to happen.
2.	$0 \leqslant P(E) \leqslant 1; E \epsilon \Omega$	
3.	$P(\cup E_i) = \Sigma P(E_i);$	for any set of disjoint E_i. This is the Axiom of Countable Additivity.

The following theorems are easy to prove.

1. $P(E) = 1 - P(E^*);$
   E^* is defined by $\Omega = E \cup E^*$, E^* and E disjoint.

2. $P(0) = 0; 0 \equiv$ null set.

3. If $E_1, E_2 \epsilon \Omega$, and $E_1 \subset E_2$, then $P(E_1) \leqslant P(E_2)$.

4. $E_1, E_2 \epsilon \Omega$, not necessarily disjoint,
   $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

Proof of 4:
$$\left.\begin{aligned} E_1 \cup E_2 &= E_1 \cup (E_1^* \cap E_2) \\ E_2 &= (E_1 \cap E_2) \cup (E_1^* \cap E_2) \end{aligned}\right\} \text{ unions of disjoint sets}$$

$$=> P(E_1 \cup E_2) = P(E_1) + P(E_1^* \cap E_2)$$

$$\text{and } P(E_2) = P(E_1 \cap E_2) + P(E_1^* \cap E_2)$$

$$\text{Hence } P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

If the outcome of an experiment can be expressed by a single number X (i.e., the outcome is either a number directly or can have a number assigned to it), then X is a random variable (r.v.). “Random” means only that the information available to us is inadequate to predict the exact outcome. “Random” does not necessarily mean “with equal probability,” contrary to its use in everyday speech. We will still sometimes use “*at* Random” to mean “with equal probability”.

If $A \subseteq \Omega$, then
$P(A) = P(X \epsilon A)$, X a r.v.

If the outcome of an experiment is an ordered set of numbers X_i, (e.g., a vector), then

$$P(A) = P[(X_1, ...X_n) \epsilon A].$$

We will use capital letters, X,Y, ... to refer to r.v.'s, lower case letters x, y, ... to refer to particular values of those r.v.'s. Thus we can speak of, e.g.

$P(X \leqslant x)$, etc.

Restriction to Real Numbers

We will restrict ourselves to two types of r.v.'s encountered in problems in the natural sciences.

1) Discrete type. The sample space Ω consists of a set of discrete points, which may be countably infinite.

Examples: Throw of dice
Sex of child (M = 1, F = 2 or whatever)
Ages of a set of people
Serial numbers of German tanks in N. Africa
Names of people (Frances = 762, George = 1001, etc.)

2) Continuous type.

Examples: Distribution of mass within a body
Momentum-transfer distribution

Thus, for both types, Ω = segments of, or points on, the real number line.

Sampling

We will always be dealing with the outcomes of experiments. These outcomes consist of a finite number of values. These values represent the *sampling* which has been made of the parent population, which may be continuous. The parent population represents the set of *values available.* The problem confronting the scientist is how to draw inferences about the parent population from the finite set of measurements in the sample.

Probability Density Function

In the discrete case, we can associate a function f(x) to each point, s.t.

$$\sum f(x) = 1 ;$$

$$P(A) = P(X \epsilon A) = \sum_A f(x).$$

f(x) is called the *probability density function* (p.d.f). In the discrete case, the p.d.f. actually represents a true probability rather than a density.

In the continuous case, we can associate a function f(x) with the probability distribution s.t. f(x) has at most a finite number of discontinuities in any finite interval and

$$f(x)\,dx = P(x \leqslant X \leqslant x + dx), \text{ with}$$

$$\int_{\Omega} f(x)\,dx = 1. \text{ (This is a Riemann integral.)}$$

Hence, f(x) is also a p.d.f. Here, the term "density" has significance, since we must multiply by dx to obtain a probability. The p.d.f. must be $\geqslant 0$ everywhere within Ω.

The normalization to 1.0, a consequence of the first axiom of probability, must always be satisfied for both discrete and continuous cases. Note that although the word "density" suggests its application to the continuous case, we use the term p.d.f. to refer to the discrete case as well.

If we know the p.d.f., we can completely describe the probability of any outcome or range of outcomes. In specifying a p.d.f. for a r.v., it is important that the range Ω be given or clearly understood.

Examples:

1) $\Omega = \{x;\, x = 0, 1, 2, 3, 4\}$

$$f(x) = \frac{4!}{x!(4-x)!}\left(\frac{1}{2}\right)^4, \quad x \epsilon\, \Omega\,; \quad \text{Note } 0! \equiv 1$$

$$P(\Omega) = \sum_{\Omega} f(x) = 1$$

Now, let $A = \{x;\, x = 0, 1\}$ be a subset of Ω; then

$$P(X \epsilon A) = \frac{4!}{0!4!}\left(\frac{1}{2}\right)^4 + \frac{4!}{1!3!}\left(\frac{1}{2}\right)^4 = \frac{5}{16}$$

2) $\Omega = \{x;\, 0 < x < \infty\}$

$$f(x) = e^{-x}; \quad x \epsilon\, \Omega \quad \text{then} \quad P(X \epsilon A) = \int_A e^{-x} dx\ .$$

Now, let $A = \{x;\, 0 < x < 1\}$, then $P(X \epsilon A) = \int_0^1 e^{-x} dx = 1 - e^{-1}$.

Cumulative Distribution Function

For the special interval covering all allowed values below (but including) a specified value x, the total probability of that interval is expressed by the *cumulative distribution function* (c.d.f). We will use upper case letters to refer to the c.d.f.

Remember, x is a real number. Let W be a r.v.

Discrete case:

$$F(x) = \sum_{w \leqslant x} f(w)$$

Continuous case:

$$F(x) = \int_{-\infty}^{x} f(w)dw$$

This is the cumulative probability for w to lie $\leqslant x$. Note that the interval specifically includes x. Note that $F(\infty) = 1$, $F(-\infty) = 0$.

Then in the continuous case $F'(x) = f(x)$ [at each point of continuity of f(x)]. In the discrete case, we have the analogue of the derivative — the Radon-Nikodym derivative with respect to the counting measure — which recovers this same relationship, but it's not particularly useful.

Properties of the c.d.f.:

$0 \leqslant F(x) \leqslant 1$
F(x) is non decreasing.
$P(a \leqslant X \leqslant b) = F(b) - F(a)$

If x is a discontinuity point of F(x), then

P(X=x) is equal to the size of the jump.

If x is a point of continuity of F(x), then

P(X=x) is 0.

Example: Discrete case

f(x)	= x/6,	x = 1, 2, 3 only;
F(x)	= 0	$x < 1$
	= 1/6	$1 \leqslant x < 2$
	= 3/6	$2 \leqslant x < 3$
	= 1	$3 \leqslant x$

In many ways, the c.d.f. is more fundamental than the p.d.f., because it refers to an actual probability rather than a probability density. Nevertheless, most applications demand the p.d.f., which is sometimes most easily derived by deriving the c.d.f. first.

Expectation Values

Consider some single-valued function u(x) of the random variable X. "u" is just some function we are interested in; it is not a p.d.f. (necessarily), or any other special function. Then the *expectation value* of u(x) is defined as

$$E[u(X)] = \sum_{x} u(x) f(x)$$

or

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)\, f(x)\, dx$$

Example:

A gambler realizes winnings u(x) upon outcome x (u(x) can be negative). Then his expected winnings are the winnings for each possible outcome x times the probability of that outcome, summed over all possible outcomes.

Properties of Expectation values.

a) If k is a constant, $E(k) = k$

b) If k is a constant, v a function, then
$$E(kv) = kE(v)$$

c) If k_1, k_2 are constants, v_1, v_2 functions, then
$$E(k_1v_1 + k_2v_2) = k_1E(v_1) + k_2E(v_2).$$
i.e., E is a "linear operator".

Sometimes we write, for the continuous case,

$$E(x) = \int x \, dF(x)$$
since $f(x)dx = dF(x)$.

Sometimes these integrals or sums don't exist, i.e., they don't converge to a finite value. We will see some cases.

Moments

Moments are certain special expectation values.

By analogy with mechanics, (for example, moments of inertia), the mth moment is given by

$$E(x^m) = \int_{-\infty}^{\infty} x^m f(x)dx$$

or

$$\sum_x x^m f(x)$$

= the mth moment of x, or the mth moment of the distribution.

The mth moment is said to exist if it is finite. Again, it doesn't always exist.

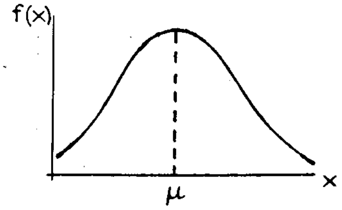
The most commonly used moment is the *mean*

$$E(X) \equiv \mu = \int_{-\infty}^{\infty} x \, f(x)dx.$$

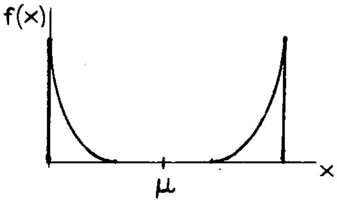
From now on, we will specifically mention the discrete case only if it somehow requires special handling.

The mean is used as a general measure of location, because it often tells roughly where most of the probability is located. In a statistical approach to estimation of the properties of an underlying distribution, the sample mean can be used as an estimate of the true population mean.

Examples:



μ useful as a measure of location



μ not especially useful as a measure of location

Central Moments

The mth central moment is given by

$$E[(x-\mu)^m] = \int_{-\infty}^{\infty} (x-\mu)^m f(x)dx, \text{ if it exists (i.e., is finite).}$$

The central moment is the moment about the mean. For a symmetrical distribution, all odd central moments are zero. For any distribution with a finite mean, the first central moment is zero.

Some of the most useful:

$E[(x-\mu)^2] \equiv$ *Variance* $\equiv \sigma^2$. Clearly, $\sigma^2 \geqslant 0$, since $f(x) \geqslant 0$ everywhere. The square root of the variance, σ, is called the *standard deviation,* and is often used as a measure of the spread of the distribution about the mean. Just as in the case of the mean, there are some distributions for which it is not a particularly apt descriptor of the shape.

Note that:

$$E[(x-\mu)^2] = E[x^2-2x\mu+\mu^2]$$

$$= E(x^2) - 2\mu E(x) + \mu^2 \quad \text{since E is a linear operator}$$

$$= E(x^2) - 2\mu\cdot\mu + \mu^2$$

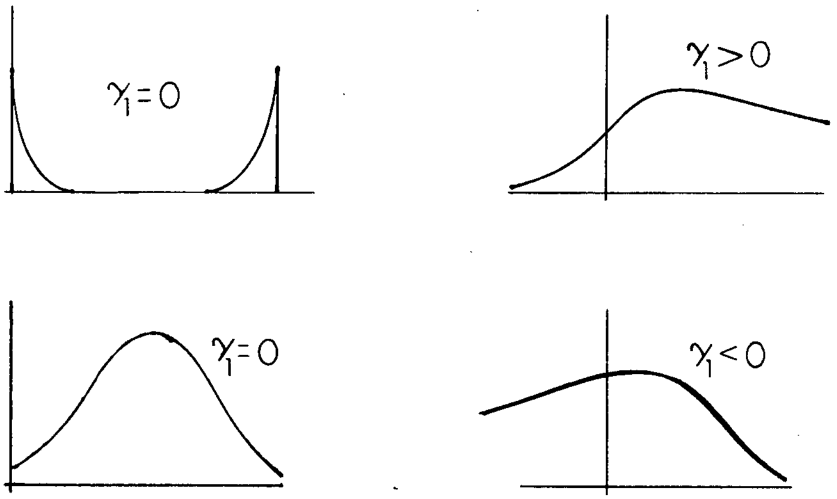
$$= E(x^2) - \mu^2$$

$$[= E(x^2) - [E(x)]^2] \ .$$

Since all symmetrical distributions have all odd central moments = 0, we can get a measure of asymmetry by looking at the third central moment, the lowest order odd moment (excluding the first): $E[(x-\mu)^3] = \int(x-\mu)^3 f(x)dx$. This has dimensions which depend upon the units employed. A more useful moment is *skewness,* defined as

$$\gamma_1 \equiv \frac{E[(x-\mu)^3]}{\sigma^3} \ ; \qquad \text{Note: } \sigma^3 = (\sigma^2)^{3/2}$$

Examples:



A measure of sharpness of peaking is given by *kurtosis*

$$\gamma_2 \equiv \frac{E[(x-\mu)^4]}{\sigma^4} - 3\,; \qquad \text{Note: } \sigma^4 = (\sigma^2)^2$$

The "-3" makes it 0.0 for a Gaussian. Therefore, $\gamma_2 > 0$ => more peaked than a Gaussian, and $\gamma_2 < 0$ => less peaked than a Gaussian.

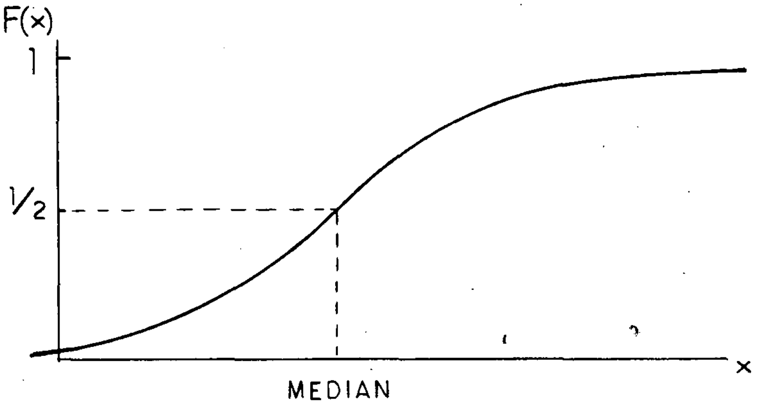
It can be shown that, if all the central moments and the mean exist, the distribution can be completely characterized by them, i.e., we can reconstruct the complete p.d.f. if we just know all the moments.

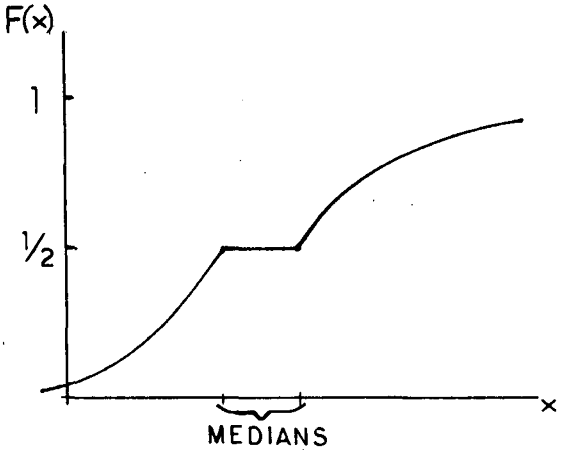
Other General Attributes of a p.d.f.

Mode — peak location (a p.d.f. can be multi-modal)

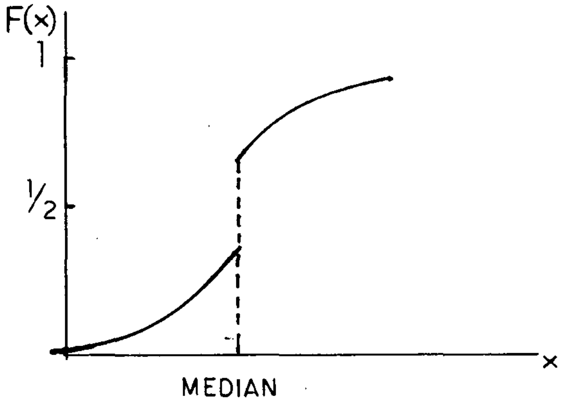
Median — defined as that point x s.t. the c.d.f. F(x) = 1/2, i.e., 1/2 of the probability lies above and below x.

$\Bigg\}$ More Measures of Location





Any number in the continuous range here may be used as the "median".



This distribution is part continuous, part discrete.

If the distribution is discrete, the median may lie between two of the allowed values. In this case, many people put it rather arbitrarily mid-way between them.

Useful Things to Remember

$$E(aX) = \int axf(x)dx = aE(X) \qquad \text{"a" a constant, X a r.v.}$$

$$\begin{aligned} V(aX) &= E[(aX)] - [E(aX)]^2 \\ &= E(a^2X^2) - [aE(X)]^2 \\ &= a^2E(X^2) - a^2[E(X)]^2 \\ &= a^2V(X) \end{aligned}$$

$$\begin{aligned} V(X+Y) &= E[(Z-\mu_z)^2]; \qquad Z \equiv X+Y \\ &= E[(X+Y-\mu_x-\mu_y)^2] = E[((X-\mu_x)+(Y-\mu_y))^2] \\ &= E[(X-\mu_x)^2] + E[(Y-\mu_y)^2] + 2E[(X-\mu_x)(Y-\mu_y)] \\ &= V(X) + V(Y) + 2E[(X-\mu_x)(Y-\mu_y)] \end{aligned}$$

The last term, called the *covariance*, has a special significance. It will be discussed in the next chapter.

The Chebychev Inequality

If μ is the mean of the p.d.f. for X, k is some arbitrarily specified positive number, and σ is the standard deviation of X (μ and σ both assumed to exist), then

$$\text{Prob}[|X-\mu| \geq k\sigma] \leq \frac{1}{k^2}\ .$$

This gives a limit to the probability that X will differ from μ by more than a specified number of standard deviations. This is true for any p.d.f. whatsoever, satisfying the conditions. Often this limit is too conservative to be really useful in specific cases.

Proof:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

$$= \int'_{|x-\mu|\leq k\sigma} (x-\mu)^2 f(x)dx + \int_{|x-\mu|\geq k\sigma} (x-\mu)^2 f(x)dx$$

break the $\int$ up into "inside" and "outside" the region.

$$\geq \int_{|x-\mu|\geq k\sigma} (x-\mu)^2 f(x)dx$$

delete the non-negative "inside" part.

$$\geq \int_{|x-\mu|\geq k\sigma} k^2\sigma^2 f(x)dx$$

replace $(x-\mu)^2$ by its minimum.

$$= k^2\sigma^2 \int_{|x-\mu|\geq k\sigma} f(x)dx = k^2\sigma^2 P[|x-\mu|\geq k\sigma]\ .$$

Therefore

$$P[|x-\mu|\geq k\sigma] \leq 1/k^2\ .$$

There is more discussion of the Chebychev Inequality in Chapter 3, by way of introduction to the normal distribution.

Chapter 2. More on Probability

Conditional Probability

Let us assume we can restrict our sample by eliminating from consideration any events not belonging to some subset $A \subseteq \Omega$. Then the probability of any event in A is increased because there are some events (those in A*, the complement of A) which are excluded (unless A* is empty). In other words, the integrated probability of all events in A is now normalized to 1.0; we wish to ignore the rest of Ω, if any. We speak of the probability of an event E, given A; we write this as $P(E|A)$. A is some restriction on Ω; it usually represents the restriction due to some fact which we believe to be true. For example: we now believe the murderer to be a man, eliminating women and children; we restrict our search for glueballs to all events containing a kaon; etc.

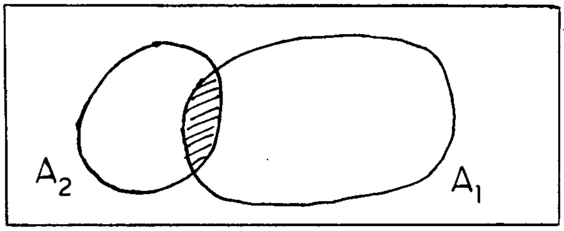
We wish this symbol to have the properties:

$$P(A|A) = 1$$

$$P(A_2|A_1) = P(A_1 \cap A_2|A_1) \ .$$

Probabilities change when we restrict the probability space, but the ratios of probabilities within the accepted set (satisfying the conditions) should not change. In particular, we require

$$\frac{P(A_1 \cap A_2|A_1)}{P(A_1|A_1)} = \frac{P(A_1 \cap A_2)}{P(A_1)}$$



$P(A_2|A_1)$ is the probability of the shaded area.

Imposing these conditions, we come to a suitable *definition*:

$$P(A_2|A_1) \equiv \frac{P(A_1 \cap A_2)}{P(A_1)} \ . \qquad \text{This assumes } P(A_1) > 0 \ .$$

It can be shown that $P(A_2|A_1)$ satisfies the axioms of probability.

Now we have the *multiplication rule*

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$$

The Joint p.d.f.

For two or more variables X_1, X_2, the *joint p.d.f.* is written $f(x_1,x_2)$.

Suppose now we want $P(a < X_1 < b)$, and we don't care what value X_2 has. Then we want $P(a < X_1 < b, -\infty < X_2 < \infty) =$

$$\int_a^b \int_{-\infty}^{\infty} f(x_1,x_2)dx_2dx_1$$

or

$$\sum_{a<x_1<b} \sum_{x_2} f(x_1,x_2)$$

The Marginal p.d.f.

This leads to the definition of the *marginal p.d.f.*

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1,x_2)dx_2$$ This is the distribution of x_1, with all the x_2-dependence integrated out.

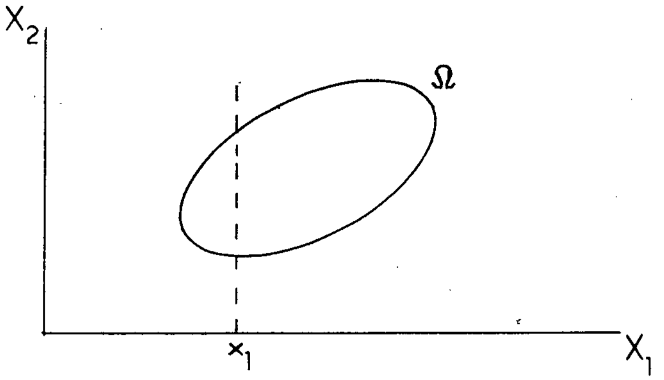
similarly,

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1,x_2)dx_1$$

The Conditional p.d.f.

Now we are ready to look at the *conditional p.d.f.*

Let X_1, X_2 be two r.v.'s on the real axis.



e.g., a Dalitz plot

Choose x_1, (a particular value of X_1) as shown. This restricts the range of x_2.

If $f_1(x_1) > 0$, then (discrete case first):

$$f(x_2|x_1) \equiv P[\underbrace{X_2=x_2}_{E_2}|\underbrace{X_1=x_1}_{E_1}] = \frac{P(E_2 \cap E_1)}{P(E_1)}$$ (conditional probability)

$$= \frac{P[X_2=x_2,X_1=x_1]}{P(X_1=x_1)} = \frac{f(x_1,x_2)}{f_1(x_1)} \, .$$

This motivates us to define a conditional p.d.f. for the continuous case in the same way. Therefore we write

$$f(x_2|x_1) = \frac{f(x_1,x_2)}{f_1(x_1)} \; ; \quad f_1(x_1) > 0 \, .$$

≡ the conditional p.d.f. for x_2 given a fixed x_1. It can be shown that this satisfies the axioms of probability. This is a function of one r.v., x_2, in the sense that x_1 is fixed. Of course, if x_1 is changed, we obtain a different function.

$P(a < X_2 < b|X_1 = x_1) = \int_a^b f(x_2|x_1)dx_2$, etc., is sometimes written

$P(a < X_2 < b|x_1)$. We can compute expectations:

$$E[u(X_2)|x_1] = \int_{-\infty}^{\infty} u(x_2)\, f(x_2|x_1)dx_2 \quad ,$$

e.g.,

$E(X_2|x_1)$ is the mean,

$E\{[X_2 - E(X_2|x_1)]^2|x_1\}$ is the variance, both given $X_1 = x_1$.

These are called the "conditional mean" and "conditional variance", respectively.

For multiple variables, we can generalize, as for example:

$$f(x_2,x_4,x_6|x_1,x_3,x_5) = \frac{f(x_1,x_2,x_3,x_4,x_5,x_6)}{f_{135}(x_1,x_3,x_5)} \quad ,$$

where

$$f_{135}(x_1,x_3,x_5) = \int\int\int f(x_1..x_6)dx_2dx_4dx_6 > 0 \, .$$

Correlation

Assume X, Y, Z are random variables with joint p.d.f. f(x,y,z). Then, e.g.,

$$\mu_x = E(X) = \int\int\int_{\Omega} xf(x,y,z)dxdydz$$

$$\sigma_x^2 = E[(X-\mu_x)^2] \, .$$

We can obtain a useful measure of the dependence of, for example, X on Y if we write

$$E[(X-\mu_x)(Y-\mu_y)] = E(XY-\mu_yX-\mu_xY+\mu_x\mu_y)$$

$$= E(XY) - \mu_x\mu_y$$

This is called the *covariance* of X and Y, COV(X,Y). Thus,

$$V(X+Y) = V(X) + V(Y) + 2E[(X-\mu_y)(Y-\mu_y)]$$

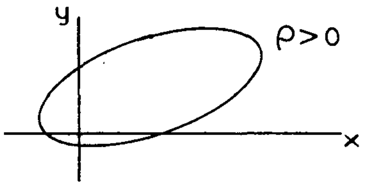
$$= V(X) + V(Y) + 2COV(X,Y)$$

We can construct a dimensionless quantity analogous to the covariance by defining the *correlation coefficient*

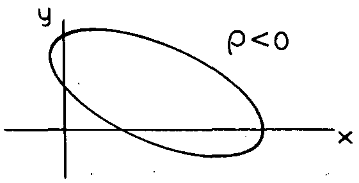
$$\rho_{xy} = \frac{COV(X,Y)}{\sigma_x \sigma_y} \ ,$$

assuming σ_x, $\sigma_y > 0$. This is scale invariant.

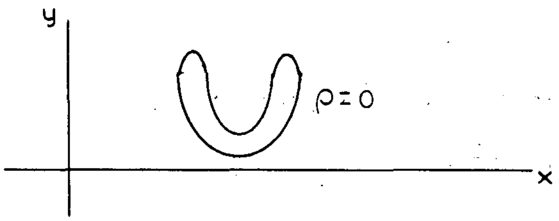
For example



contour of constant p.d.f., e.g., boundary of region p.d.f. > 0.



ρ is a measure (an imperfect one) of how well Y and X depend upon on one another. It is most useful when the contours of constant p.d.f. look as above, more or less elliptical.



X and Y are clearly related in probability. Here, ρ is not very revealing as a characterization of the variation of X with Y. We could always define higher-order correlation terms, but this is not often done.

To prove this last case, we note that if

$$E[(X-\mu_x)|y] = \int \frac{(x-\mu_x)f(x,y)dx}{f_2(y)} = 0 \text{ for all y (i.e., the mean X is independent of y),}$$

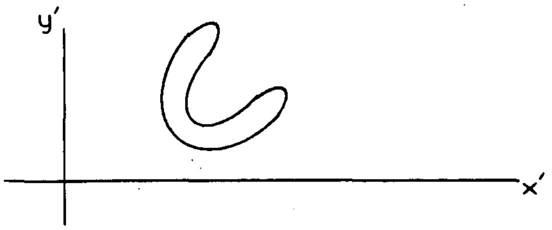
then since

$$COV(X,Y) = \int (y-\mu_y) \int (x-\mu_x) f(x,y)dx\, dy, \text{ and the above gives us}$$

$$\int (x-\mu_x)f(x,y)dx = 0 \text{ for all y, we establish}$$

$$COV(X,Y) = 0\ .$$

We could *change variables* to a correlated set:



$\rho \neq 0.$

The zero correlation coefficient in the previous case was an "accidental" consequence of the genuine symmetry of the p.d.f. and the alignment of the axes.

We can establish that $-1 \leqslant \rho \leqslant 1$:

Let $U = X - \mu_x$; $V = Y - \mu_y$; choose some constant, k.

Then

$$E[(U-kV)^2] = k^2E(V^2) - 2kE(UV) + E(U^2) \ .$$

This is a quadratic in k, which has zeroes only if

$$4[E(UV)]^2 - 4E(V^2)E(U^2) \geqslant 0 \, .$$

But $E[(U - kV)^2] \geqslant 0$ for all k, i.e., it cannot pass *through* zero; therefore any roots are degenerate and occur at $[E(UV)]^2 = E(V^2)E(U^2)$. Everywhere else, $[E(UV)]^2 - E(V^2)E(U^2) < 0$. Since $E(UV) = COV(X,Y)$, $E(U^2) = \sigma_X^2$ and $E(V^2) = \sigma_Y^2$, the theorem is proved.

The case $\rho = \pm 1$ occurs if and only if $E[(U-kV)^2] = 0$ for some k, and therefore $U = kV$ everywhere with probability 1.0; thus $X - \mu_x = k(Y - \mu_y)$ is just a straight line. The slope k cannot be 0 or infinite unless all of the probability of U or V is concentrated at a single point.

The correlation ρ is a measure of the *average* linear change in the marginal p.d.f. of one variable for a specified change in another variable. In the upright horseshoe-shaped case we saw earlier, the correlation is zero, but the variables clearly depend upon one another. That is, if we make some small change in one variable, the marginal p.d.f. for the other will change, but only in shape; the change in the average is zero. Another concept is needed to discuss this sort of variation.

Dependence and Independence

We know that

$$f(x_1,x_2) = f(x_2|x_1)f_1(x_1).$$

Suppose $f(x_2|x_1)$ has no dependence on x_1. Then

$$\begin{aligned} f_2(x_2) &= \int f(x_2|x_1)f_1(x_1)dx_1 \\ &= f(x_2|x_1)\int f_1(x_1)dx_1 \\ &= f(x_2|x_1), \qquad \text{by the normalization of } f_1(x_1). \end{aligned}$$

Therefore, $f(x_1,x_2) = f_1(x_1)f_2(x_2)$, the product of the marginals.

Definition

The r.v.'s X_1 and X_2 are *independent* if

$$f(x_1,x_2) = f_1(x_1)f_2(x_2).$$

Otherwise, they are dependent.

In order to be independent, the domains Ω_1 and Ω_2 also have to be independent. In the horseshoe-shaped example above, the variables are dependent, even though their linear correlation is zero.

Theorem

Take two r.v.'s X_1, X_2 with joint p.d.f. $f(x_1,x_2)$. Then X_1, X_2 are independent if and only if we can find functions g and h s.t.

$f(x_1,x_2) = g(x_1)h(x_2)$, where

$g(x_1) > 0$ for $x_1 \epsilon A_1$; $h(x_2) > 0$, $x_2 \epsilon A_2$, zero elsewhere.

The proofs of this theorem and the next are left as an exercise.
Hint: We can show $g(x_1) = c_1f_1(x_1)$; $h(x_2) = c_2f_2(x_2)$, with c_1, c_2 constants such that $c_1c_2 = 1$.

Theorem

Let X_1 and X_2 be independent with marginal p.d.f.'s $f_1(x_1)$ and $f_2(x_2)$, respectively. Then

$E[u(X_1)v(X_2)] = E[u(X_1)]E[v(X_2)]$,

u and v functions, provided all these expectations exist.

A consequence of these theorems is that: If X_1 and X_2 are independent,

$E[(X_1-\mu_1)(X_2-\mu_2)] = E(X_1-\mu_1)E(X_2-\mu_2)$ and therefore

$COV(X,Y) = 0$. Hence

$$V(X+Y) = V(X) + V(Y) + 2COV(X,Y)$$

$$= V(X) + V(Y) \text{ if (not only if) independent.}$$

Bayes' Theorem

First, the law of total probability. Let B_i be a set of mutually exclusive subsets of Ω, and let $\sum_{i=1}^{n} P(B_i) = 1$. Then the B_i are said to be exhaustive, i.e., all of the probability is covered by the B_i. Now take some $A \epsilon \Omega$. "A" may or may not be one of the B_i. Then the law of total probability states

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) .$$

That is, partition A up into the pieces in each B_i. Then the total probability of A is the sum of the probabilities of the pieces. Technically, the B_i only have to cover the whole of Ω which has non-zero probability. Pieces which have zero probability, e.g., isolated points for a continuous p.d.f., can be excluded.

This law follows from the multiplication rule (Chapter 1).

We saw earlier that $f(x_2|x_1)$ was a p.d.f. for x_2, in the sense that x_1 is fixed. But if we change x_1, we get a change in $f(x_2|x_1)$ (unless X_1 and X_2 are independent). We will be seeing how, if we already know x_2, we can turn this around and make probabilistic statements about x_1. Bayes' Theorem is one approach.

Theorem

Let B_i be a set of mutually exclusive and exhaustive events. Take any event A s.t. $P(A) > 0$. Then

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{P(A)}$$

$$= P(A|B_i)\frac{P(B_i)}{\sum_i P(A|B_j)P(B_j)} ; \quad i = 1, ..., n$$

This shows us how to reverse the order of the statement in a conditional probability. The first part follows from

$$P(B_i \cap A) = P(A|B_i)P(B_i) = P(B_i|A)P(A)$$

(the part of A that is in B_i is the same as the part of B_i that is in A). The second part follows from the law of total probability.

In our use of Bayes' Theorem, we normally have a situation in which A represents a certain set of data. We want to evaluate the respective probabilities of a certain set of explanations or models B_i, in order to select one of them, or at least eliminate some as unlikely. In other words, given the outcome A of our experiments, can we say anything about the different models or theories represented by B_i? The B_i are most frequently some set of parameters whose values we desire. Since we can in principle evaluate the probabilities of the A for any assumed B_i, $P(A|B_i)$, the job of the statistician would be done if we only knew what to use for $P(B_i)$.

A simple example:
1/1000 adults has a certain disease (we restrict ourselves, if necessary, to some subsample of adults where the probability is constant and known). A certain diagnostic test has the following properties:

— if the person has the disease, the test gets a positive result 99% of the time.

— if the person doesn't have the disease, it gets a positive result 2% of the time.
You test positive — what is the probability that you have it?

Data:	$A \equiv$ positive test	
Models:	$B_1 \equiv D$ (diseased) $B_2 \equiv D^*$	exhaustive and exclusive
Want:	$P(B_1\|A)$	
Have:	$P(B_1) = 1/1000$ $P(B_2) = 999/1000$	general statements made *before* the data is taken
	This information gives us our *prior* distribution, which is $P(B_i)$.	
We also know:	$P(A\|B_1) = .99$ $P(A\|B_2) = .02$	These don't have to add to 1, since each is just the fraction of B_i covered by A. Their sum is bounded from above only by $P(B_1\|B_1) + P(B_2\|B_2) = 2$.

We can now compute $P(A) = 0.990 \times 0.001 + 0.020 \times 0.999 = 0.02097$, from the law of total probability:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

Then,

$$\left.\begin{aligned} P(B_1|A) &= \frac{.99 \times .001}{.99 \times .001 + .02 \times .999} = .047 \\ P(B_2|A) &= \frac{.02 \times .99}{.99 \times .001 + .02 \times .999} = .953 \end{aligned}\right\}$$

These do add to 1, since the B_i are exhaustive and exclusive, and together they cover A exactly.

This is our posterior distribution. The data have increased our estimate of the probability that you have the disease from .001 to .047. It's still small. Why? Because so many more people don't have the disease, the small chance that they will test positive anyway leads to more than 20 of them testing positive for each one that actually has the disease (and tests positive).

This can be calculated from a tree diagram without referring to Bayes' Theorem. But most cases of interest to us will have continuous variables, or a large number of discrete variables, so a tree diagram is not convenient.

Another example: in the case of ambiguity between Cabibbo-favored and unfavored decays of observed charmed particle candidates, a certain experiment chooses to report the favored decays. The unfavored decays are *a priori* considered to be unlikely.

We use Bayes' Theorem to modify prior beliefs by incorporating the information of the data to obtain better educated beliefs. Frequently, as in these examples, our prior beliefs are based upon an understanding of the general population and we are interested in using that plus some measurements to get the best possible understanding of some particular case.

We will normally use this to help us find θ, some parameter of nature. Thus, we treat the state of nature θ as a random variable to which we assign a prior p.d.f. $P(\theta)$ if we can. If we then do an experiment which has outcome z, and if we know the p.d.f. of any z as a function of θ, $f(z|\theta)$, then we can get the p.d.f. of z,

$$f_1(z) = \int f(z|\theta)P(\theta)d\theta$$

or

$$f_1(z) = \sum_i f(z|\theta_i)P(\theta_i)\,.$$

Since we treat θ as a r.v., it makes sense to talk about marginal p.d.f.'s

Then we can improve our ideas about the state of nature θ by

$$\underset{\substack{\uparrow\\ \text{posterior}}}{h(\theta|z)} = \frac{f(z|\theta)P(\theta)^{\text{prior}}}{f_1(z)}\quad.$$

In the continuous case, P is a p.d.f., not a probability.

Bayes' Theorem is mathematically impeccable. However, there is much controversy about the choice of prior, in some cases. Most workers in the sciences feel it is best to stick to safe cases, where the prior is really understood.

It is said that more statistical literature is published about the choice of prior than about any other subject. If the prior is unknown, what can we do? The good Reverend Bayes is responsible for:

Bayes' Postulate. If we are completely ignorant about θ, express this ignorance by setting $P(\theta) = 1$. So each θ is *a priori* equally likely. Notice, first of all, and least important, that this is what is called an *improper* prior, because $\int_\Omega P(\theta) \neq 1$. That turns out to be okay, anyway. What is not okay to most scientists is the attempt to get something for nothing. If we don't know anything about θ, how can we turn around and say $P(\theta) = 1$? That is information which gives us a different posterior $h(\theta|z)$ than some other choice of $P(\theta)$. But what if we decided we were completely ignorant about $\theta^{3/2}$, or $1/\theta$, or (θ^2 + the speed of the current on the west coast)? Are we more ignorant or less ignorant about these? A little ignorance can be an arbitrary thing.

In these notes we will express our ignorance by ignoring Bayes' Postulate. But watch for its head to appear when we discuss Maximum Likelihood.

In some cases, the complete

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

cannot be found because, although we know some of the terms, we don't know them all. In this case we can still find *relative* probabilities for $P(B_i|A)$:

$$P(B_i|A) \propto P(A|B_i)P(B_i)$$

or

$$h(\theta|z) \propto f(z|\theta)g(\theta)\,.$$

Change of Variables

Let X be a discrete r.v., with p.d.f. f(x), $X \epsilon A$. Let $y = u(x)$ define a one-to-one transformation $A \rightarrow B$, $y \epsilon B$. We want to find the p.d.f. of y. Find the inverse transformation $x = w(y)$ which maps B back onto A. Then the events

$$Y = y \quad : \text{Y assumes the value y ;}$$

and

$$X = w(y) \quad : \text{X assumes the value w(y) ;}$$

have the same probability. Therefore

$$g(y) \equiv P[Y{=}y] = P[X{=}w(y)] = f[w(y)], \; y \epsilon B, \text{ discrete.}$$
$$= 0 \qquad \text{elsewhere}$$

Example: X has the p.d.f.

$$f(x) = \frac{3!}{x!(3-x)!}(\frac{2}{3})^x(\frac{1}{3})^{3-x}, \; x = 0,1,2,3 \;;$$

we want $Y = X^2$. Such a transformation is not usually one-to-one, but here, there is no negative x, so it's okay; the inverse is $X = \sqrt{Y}$, and not $-\sqrt{Y}$.

$$g(y) = \frac{3!}{\sqrt{y}!(3-\sqrt{y})!}(\frac{2}{3})^{\sqrt{y}}(\frac{1}{3})^{3-\sqrt{y}}, y = 0,1,4,9.$$

In the case of joint p.d.f.'s of two or more r.v.'s, we must have as many variables in our new set as there were in our old set. Later we can reduce the number, if desired, by obtaining the marginal p.d.f.

Example:

Take two r.v.'s, X and Y, with

$$f(x,y) = \frac{\mu_x^x \mu_y^y e^{-\mu_x} e^{-\mu_y}}{x!y!} \quad ; x,y = 0,1,2,3... \text{ (joint Poisson)}$$
$$= \qquad 0 \qquad \text{elsewhere}$$

We want the p.d.f. of $Z = X+Y$, a single variable. We must have a second variable as well, because we started with two. Choose something simple, e.g., Y itself.

$$Z_1 = X + Y$$

$$Z_2 = Y$$

Z_1 and Z_2 must be different; they also must not be functions of each other; i.e., if you know one, you cannot obtain the other without knowing something else such as the values of X or Y.

Then

$B = \{(z_1,z_2), z_1 = 0,1,2, \ldots; z_2 = 0,1, \ldots, z_1\}$; $0 \leqslant z_2 \leqslant z_1$

The inverse functions are

$X = Z_1 - Z_2$; $Y = Z_2$.

These are unique.

Then

$$g(z_1,z_2) = \frac{\mu_x^{z_1-z_2}\mu_y^{z_2}e^{-\mu_x-\mu_y}}{(z_1-z_2)!\,z_2!} \quad , \quad (z_1,z_2) \epsilon B .$$

This gives us the joint p.d.f. What about Z_1?

$$g_1(z_1) = \sum_{z_2=0}^{z_1} g(z_1,z_2)$$

$$= \frac{e^{-\mu_x-\mu_y}}{z_1!} \sum_{z_2=0}^{z_1} \frac{z_1!}{(z_1-z_2)!z_2!} \mu_x^{z_1-z_2}\mu_y^{z_2}$$

$$= \frac{(\mu_x+\mu_y)^{z_1} e^{-\mu_x-\mu_y}}{z_1!} \quad , \quad z_1 = 0,1,2 \cdots \quad \text{(by the Binomial Theorem)}$$

$$= 0 \qquad\qquad \text{elsewhere}$$

This is a Poisson also, as we will see when we study the Poisson.

Example — Continuous Case

Let $f(x) = 2x, \qquad 0 < x < 1$

$= 0 \qquad$ elsewhere

Define $Y = 8X^3$ and find g(y). This transformation is one-to-one. Go back to basic probability. For all $0 < a < b < \infty$, the event $a < Y < b$ will occur if and only if

$\frac{1}{2}\,{}^3\sqrt{a} < X < \frac{1}{2}\,{}^3\sqrt{b}$, since $Y = a$ implies that $X = \frac{1}{2}\,{}^3\sqrt{a}$, etc. Therefore

$$P[a < Y < b] = P[\frac{1}{2}\,{}^3\sqrt{a} < X < \frac{1}{2}\,{}^3\sqrt{b}]$$

$$= \int_{{}^3\sqrt{a/2}}^{{}^3\sqrt{b/2}} 2x dx = G(b) - G(a),$$ where G is the c.d.f. of Y, by the definition of the c.d.f.

We can derive the p.d.f. for Y from this if we change the variable of integration to dy:

$$\frac{dx}{dy} = \frac{1}{6y^{2/3}} \;, \quad \text{from } x = \frac{1}{2}\,{}^3\sqrt{y} \;;$$

Therefore:

$dx = \frac{dy}{6y^{2/3}}$. Now we can write

$$P[a<Y<b] = \int_a^b 2\left(\frac{\sqrt[3]{y}}{2}\right)\frac{1}{6y^{2/3}}\,dy = \int_a^b \frac{1}{6y^{1/3}}dy, \text{ and}$$

$g(y) = \frac{1}{6y^{1/3}}$. We never actually had to do the integral.

The steps were:

1) find the inverse transformation $x = w(y)$. Here, $w(y) = \frac{1}{2}\sqrt[3]{y}$. Note that $x \epsilon A$ and $y \epsilon B$. Of course, both are subsets of the real axis.

2) Assume $w'(y)$ continuous and $\neq \infty$ for all $y \epsilon B$. Then $g(y) = f[w(y)]|w'(y)|$, because we must take the absolute value of w' to change variables in an integral. Because the p.d.f. is a *density,* i.e., a probability per unit length, we must do this step of multiplying by $|w'(y)|$, which is called the Jacobean, to take account of the change in the size of the unit of length.

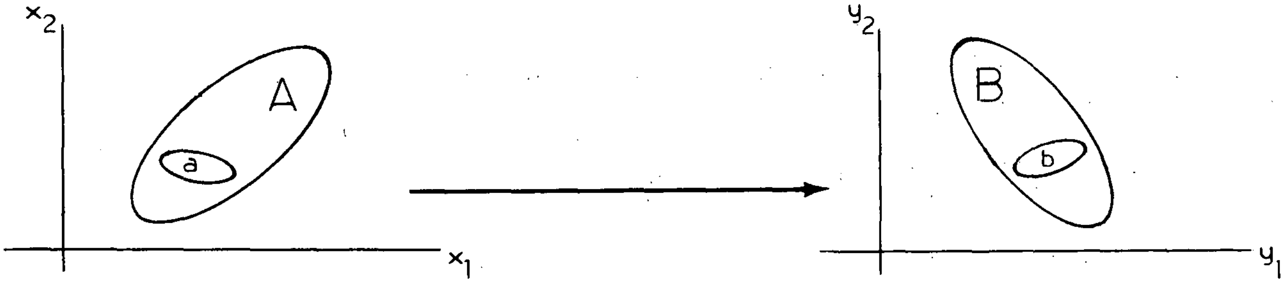
A Mnemonic: $w'(y) = \frac{dx}{dy}$, therefore think of $g(y)dy = f(x)dx$ where dy and dx are both set to be > 0, since lengths are taken positive.

We will write the Jacobean, expressing the change in the unit of length, as

$$J = \left|\frac{dx}{dy}\right| .$$

For multiple variables, we must use the multidimensional equivalent:

Look at $(X_1, X_2) \rightarrow (Y_1, Y_2)$, a one-to-one transformation.



We wish to map $A \rightarrow B$. "a" is a small subset of A which maps onto the small subset "b". That is, the events $(X_1, X_2) \epsilon a$ and $(Y_1, Y_2) \epsilon b$ are equivalent and occur with the same probability. Then

$$P[(Y_1, Y_2) \epsilon b] = P[(X_1, X_2) \epsilon a] = \iint_a f(x_1,x_2)dx_1dx_2 .$$

Now change variables of integration. First, invert: If $y_1 = u_1(x_1, x_2)$, $y_2 = u_2(x_1, x_2)$, then $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$.

The transformation must be one-to-one, so these exist and are unique. We must also assume all the first derivatives of w_1 and w_2 exist. Both of these conditions can be relaxed for certain directions of transformation in certain cases, with sufficient care.

$$\text{Then } \iint_a f(x_1, x_2)dx_1dx_2 = \iint_b f[w_1(y_1, y_2), w_2(y_1,y_2)]\underbrace{|J|dy_1dy_2}_{\text{new unit of area}}$$

$$J = \begin{vmatrix} \dfrac{\partial w_1}{\partial y_1} & \dfrac{\partial w_1}{\partial y_2} \\ \dfrac{\partial w_2}{\partial y_1} & \dfrac{\partial w_2}{\partial y_2} \end{vmatrix}$$

This is a determinant, which is non-singular because the transformation is one-to-one.

This is proved in calculus.

Then $g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|$.

There is some additional discussion about change-of-variables in Chapter 9, "Propagation of Errors."

From this theory, we can calculate the expectation of functions of random variables. Assume a r.v. X with p.d.f. f(x). We wish to find E[u(X)], where u(X) is some function of X which is continuous and has a unique inverse. Then u(X) is itself a r.v., Y = u(X), with X = w(Y). The p.d.f. of Y is g(y).

$$E[u(X)] = E(Y) \equiv \int yg(y)dy \equiv \int u(x)f(x)dx$$

Therefore we can evaluate E[u(X)] in either of two ways, whichever is more convenient. If we have Y and g(y), we can use the first form. Alternately, we may have f(x) and be able to express Y as a function of x. Then we don't need to get the p.d.f. of Y.

In the case of two or more variables, u = g(x,y), we have

$$E(u) = \iint g(x,y)f(x,y)dxdy.$$

For example,

$$\begin{aligned} E(X+Y) &= \iint (x+y)f(x,y)dxdy \\ &= \iint xf(x,y)dxdy + \iint yf(x,y)dxdy \\ &= E(X) + E(Y) \end{aligned}$$

Chapter 3. Some Special Distributions (Mostly Univariate)

Uniform

$$U(x) = f(x) = \frac{1}{a}\ ; 0 \leqslant x \leqslant a\ .$$

$$E(x) = \int_0^a \frac{x}{a}\,dx = \frac{a}{2}$$

$$V(x) = \frac{1}{a}\int_0^a x^2dx - \frac{a^2}{4} = \frac{a^2}{12}$$

Bernoulli Trials

A Bernoulli Trial has just two outcomes, e.g., the toss of a coin. We assign a r.v. k of 0 or 1 to each outcome, with probabilities as follows:

$$\text{Outcomes}\begin{cases} \underline{k} & \underline{\text{Probability}} \\ 0 & q \\ 1 & p=1-q \end{cases}$$

$f(k) = p^kq^{1-k}$,which returns the above probabilities.

The only parameter is q (or p).

Moments:

$$E(X^m) = E(X) = 1 \times p + 0 \times q = p\ .$$

$$V(X) = E(X^2) - [E(X)]^2 = p - p^2 = pq$$

Binomial

Suppose we have n independent Bernoulli trials. Define a "success" to be a result of 1. For all n to be "successes," the probability is p^n. If there are $k = n - 1$ successes, the probability is $n \times p^kq$ (there are n ways to have the "failure"). In general, if there are k successes, $0 \leqslant k \leqslant n$, P(k successes in n trials) =

$$\binom{n}{k} p^kq^{n-k}\ ;\ \binom{n}{k} = \frac{n!}{k!(n-k)!}\ .$$

We will write B(k;n,p). This is the binomial distribution for k successes out of n Bernoulli trials, with probability of success p on each trial. k is the sum of the outcomes, each of which is either 0 or 1.

$$E(k) = nE(k = 1) = np$$ since E is a linear operator and the trials are independent.

It is a trivial exercise to show that the variance of the sum of trials is

$$V(k) = \sigma_K^2 = npq = np(1 - p) \ ,$$

since the trials are all independent, and therefore uncorrelated.

Reproductive Property of the Binomial

Take X,Y independent r.v.'s distributed according to binomial p.d.f.'s with parameter q as above. Then

$$f(x,y) = \begin{pmatrix} n_x \\ x \end{pmatrix} p^x q^{(n_x - x)} \begin{pmatrix} n_y \\ y \end{pmatrix} p^y q^{(n_y - y)}$$

What is the distribution of X + Y? The change of variables technique is described in the preceding chapter. Take:

$$Z_1 = X + Y \ , \quad Z_2 = Y,$$

$$X = Z_1 - Z_2 \ ; n_{z_1} = n_x + n_y \ , n_x = n_{z_1} - n_{z_2}$$

$$Y = Z_2 \ ; n_{z_2} = n_y \ , n_y = n_{z_2}$$

$$g(z_1 \ , z_2) = \begin{pmatrix} n_{z_1} & - & n_{z_2} \\ z_1 & - & z_2 \end{pmatrix} \begin{pmatrix} n_{z_2} \\ z_2 \end{pmatrix} p^{z_1} q^{(n_{z_1} - z_1)}$$

Now the distribution for X + Y will be the marginal of this, summed over z_2 from 0 to n_{z_2}. We can see immediately from inspection (since the p and q terms come out of the sum) that this p.d.f. must have the form

$$g_1(z_1) = C(n_{z_1} \ , z_1) \ p^{z_1} q^{(n_{z_1} - z_1)} \ ,$$

where C is some coefficient which does not depend upon either p or q. Since this must be a p.d.f.,

$$C(n_{z_1} \ , z_1) = \begin{pmatrix} n_{z_1} \\ z_1 \end{pmatrix} \ ,$$

otherwise it's not properly normalized. Therefore $Z_1 = X + Y$ is also a Binomial. This is a very useful property many distributions have, called the Reproductive Property. The sum of two or more independent r.v.'s with certain p.d.f.'s is itself distributed according to that p.d.f. (usually with different parameters).

Example of use of the Binomial

Consider three independent r.v.'s X_1, X_2, X_3, each with the same p.d.f., f(x), which could be anything, discrete or continuous. Let Y be the middle value, that is, whichever of the X_i's that lies between the other two in value. Determine the p.d.f. of Y.

We begin with the c.d.f., which is easily defined. The total probability that the middle value lies below some chosen number y is equal to the probability that at least two of the X's do, simultaneously:

$$G_Y(y) \equiv P(Y \leqslant y) = P(2 \text{ or } 3 \text{ of the } X_i \leqslant y)$$

We will adopt the notation G_Y and F_X, in a few cases where confusion could arise from the arguments.

One Bernoulli trail is to see if one $X_i \leqslant y$. If so, count one success. The probability of one success is

$$p = F_X(y) = P(X_i \leqslant y)$$

for any X_i, since they all have the same F_X. Then

$$G_Y(y) = B(2\,;3\,,p = F_X(y)) + B(3\,;3\,,p)$$

$$= \binom{3}{2} p^2q + \binom{3}{3} p^3 = 3p^2q + p^3$$

$$= [F_X(y)]^2 [3 - 3F_X(y) + F_X(y)] = [F_X(y)]^2 [3 - 2F_X(y)]\ .$$

We cannot evaluate this any further until we know F_X. But we can do:

$$g(y) = \frac{dG_Y(y)}{dy} = 2F_X(y)\, f_X(y)[3 - 2F_X(y)] + [F_X(y)]^2 [-2f_X(y)]$$

$$= 6F_X(y)\, f_X(y)[1 - F_X(y)]\ .$$

Multinomial

Consider a histogram with k bins. The probability that a given event will fall in bin "i" is p_i. Then, for a fixed number of events, the distribution of events among the bins is

$$M(x_i\,;p_i\,,n) = \frac{n!}{x_1!\,x_2!\,\cdots\,x_n!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}\ ,$$

where n = number of events, x_i = contents of ith bin, and $\sum_1^k p_i = 1$. We will not write k as an explicit parameter; it is evident in the number of terms in the products. For a given bin, an event is either in it (success), or out of it (failure). Therefore the marginal for that bin is a Binomial, and

$$E(x_i) = np_i\ ,$$

$$V(x_i) = \sigma_i^2 = np_i(1 - p_i)\ .$$

This is a *multinomial* distribution.

Note that these x_i are not independent, because $x_k = n - \sum_{j=1}^{k-1} x_j$.

Normally one thinks of the distribution of the number of events in a given bin as a Poisson (below). This is indeed the case once we remove the stipulation that n is fixed in advance.

Another application is in the case of a particle or nucleus which has k − 1 different decay modes. "x_k", the number in the last bin, might represent those which do not decay in the observing time. Then M describes the number of decays into each mode.

Poisson

This is a very important model. We can look at a certain process which develops as a function of time or space. We consider discrete events, which either happen or don't. As a function of the parameter x (space or time), if we can assume

1) the number of events in any interval of x is independent of the number in any other (non-overlapping) interval;

2) in any small interval Δx, the probability of one event is proportional to Δx and the probability of two or more events vanishes as $\Delta x \to 0$. That is,

$$P_{\lambda\Delta x}(1) = \lambda\Delta x + O(\Delta x) \text{ as } \Delta x \to 0 \ ,$$

where $O(\Delta x)$ is defined by

$$\lim_{\Delta x \to 0} \frac{O(\Delta x)}{\Delta x} = 0$$

and

$$\sum_{n>1} P_{\lambda\Delta x}(n) = O(\Delta x) \text{ as } \Delta x \to 0 \ ;$$

3) This relationship is independent of x. In particular, λ does not depend upon x.

Then

$$P_{\lambda x}(n) = \frac{e^{-\lambda x}(\lambda x)^n}{n!}$$

$$\equiv \frac{e^{-\mu}\mu^n}{n!} \ ; \mu = \lambda x, n = 0,1,2 \cdots \qquad \mu \text{ is the Poisson parameter.}$$

Then

$$\sum_{n=0}^{\infty} P_\mu(n) = e^{-\mu} \sum \frac{\mu^n}{n!} = 1 \qquad \text{as needed.}$$

The Poisson is very asymmetrical for small μ, but becomes more symmetrical as μ increases. It has fairly good symmetry even by $\mu = 5$ or 6. The outcome of any experiment observing a Poisson process must be an integer. The Poisson parameter, however, may be any non-negative real number.

$$E(n) = \sum_{n=0}^{\infty} n \, \frac{\mu^n e^{-\mu}}{n!}$$

$$= e^{-\mu} \mu \sum_{n=1}^{\infty} \frac{\mu^{(n-1)}}{(n-1)!} = e^{-\mu}\mu \sum_{s=0} \frac{\mu^s}{s!} \; , \; s = n - 1$$

↑ ↑

Note the changes in the
limits of summation.

$$= \mu \; .$$

In a similar fashion, we can establish (exercise) that $V(n) = \mu$. It is the same as the mean.

We have already shown (Chapter 2) that if X and Y are Poisson, then

$$g(x + y) = \frac{(\mu_x + \mu_y)^{x+y} e^{-\mu_x - \mu_y}}{(x + y)!} \quad ,$$

which is $P_{\mu_x+\mu_y}(X + Y)$, and therefore the sum of two Poisson r.v.'s is also Poisson, and the Poisson possesses the — *Reproductive Property.*

This means that if we observe signal X and background Y in an experiment, each of which is an independent Poisson, then the sum is also Poisson. This has important applications for background subtraction, in which we must be able to estimate errors.

It can be shown (exercise) that the interval size s is distributed as $f(s) = \lambda e^{-\lambda s}$, a distribution known as the exponential, which will be covered shortly. This is the interval in time "s" from one event to the very next, or its analogue if the parameter s represents a space dimension. The derivation does not actually depend upon starting at the previous event—we could have started anywhere, and just timed to the first event after starting.

Here, we could substitute distance for time, and look for the distribution of e.g., single scattering collisions of angle greater than "something" along a track. The use of the Poisson there, of course, assumes that no process (e.g., slowing down) changes the scattering probability appreciably, since the probability must be the same in each small interval.

Poisson Approximation to a Binomial Problem

Consider the case of a collection of n radioactive atoms. In any particular time interval, T, some of them will decay and others not. Therefore we will have a binomial distribution of the number "r" which decay:

$$P(r) = B(r \, ; n \, , p) = \frac{n!}{r!(n - r)!} \, p^r q^{n-r}$$

p = some probability for one atom which depends, of course, on T. The difference between this case and the Poisson is the limit "n" to the number of atoms: as more and more decay, there are less and less left to decay. Therefore the Poisson proportionality λ depends on T, and so the number of decays per unit time interval, or in time interval T, is not Poisson. But if there are a lot of atoms, and $pn << n$, then this depletion is not noticed over the time interval of a given experiment. Then $r <<$ n with high probability, and

$$\frac{n!}{(n-r)!} = n(n-1)(n-2)...(n-r+1) \text{ r terms}$$

$\cong n^r$, since each term is about equal to n. Also,

$$q^{n-r} = (1 - p)^{n-r} = 1 - p(n - r) + \frac{p^2}{2!}(n - r)(n - r - 1) + \cdots$$

$$\cong 1 - p(n - r) + \frac{p^2}{2!}(n - r)^2 + \cdots$$

$$= e^{-p(n-r)} \cong e^{-pn} \ .$$

Therefore,

$$B(r \, ; n \, , p) \cong \frac{n^r}{r!} p^r e^{-np} = \frac{(np)^r}{r!} e^{-np} \ ,$$

which is a Poisson of parameter np.

NOTE: As soon as n changes noticeably, the Poisson approximation is no longer valid. This is equivalent to saying that we are approximately Poisson as long as we are on the flat portion of the exponential decay law.

The only difference between the true Binomial and the Poisson laws governing the number of decays in T is that the Poisson has no depletion of the source of events, so if the source is very large, and depletion is slow, the Poisson is fine.

Exponential

We can derive the exponential decay law from the binomial: Let "p" be the probability that any given atom decays in time dt (p is a constant of nature), let "r" be the actual number of decays in dt, and "n" be the total number of atoms at the moment. Then

$$P(r) = \frac{r!}{r!(n - r)!} p^r q^{n-r} \ ,$$

where n is the *current* number (at the start of dt). Now, r is just the current value of $|dn/dt|$. Therefore $E(dn/dt) = -E(r) = -np$ (the negative sign reflects, of course, that positive r decreases n), and

$$E(n) = n_0 e^{-pt} \ .$$

We shall see, from the law of large numbers, that the probability that n differs from its expectation by more than a small amount decreases as n increases. Therefore we usually say

$$n = n_0 e^{-pt} \ .$$

The true meaning of this expression is that n is intended to represent the *expected* number of events.

The *exponential* p.d.f., describing the distribution of individual decay times, is given by

$$f(t) = \frac{1}{t_o} e^{-t/t_o} . \ E(t) = t_o \, ; V(t) = t_o^2 \ .$$

As mentioned previously, the exponential also describes the distribution of interval lengths between events in a Poisson process.

Gamma

Start from the above exponential distribution of interval lengths in a process accurately (or accurately enough) described by a Poisson: Suppose we want the distribution of the time T to k events. The r.v. T is the sum of the individual t_i (the length of the i^{th} interval), each of which is exponential. Each of the t_i is independent. Choose a value t.

$$G(t) = P(T \leqslant t) = 1 - P(T > t) \ .$$

Note that the case T > t means that there are fewer than k events in t.

Now we use the property of the c.d.f. that

$$P(T > t) = \sum_{x=0}^{k-1} P(X = x) = \sum \text{all probabilities to have} < k \text{ events}$$

$$= \sum_{x=0}^{k-1} \frac{(\lambda t)^x e^{-\lambda t}}{x!} \quad , \quad \text{where } \lambda \text{ is the appropriate parameter.}$$

From calculus (mathematical induction) it can be shown that

$$\sum_{x=0}^{k-1} \frac{(\lambda t)^x e^{-\lambda t}}{x!} = \int_{\lambda t}^{\infty} \frac{z^{k-1} e^{-z}}{(k-1)!} dz$$

Therefore

$$G(t) = 1 - \int_{\lambda t}^{\infty} \frac{z^{k-1} e^{-z}}{\Gamma(k)} dz \ .$$

Γ is the gamma function: $\Gamma(k) = (k - 1)!$, for k an integer.

$$= \int_{0}^{\lambda t} \frac{z^{k-1} e^{-z}}{\Gamma(k)} dz \ .$$

If $t \leqslant 0$, $G = 0$. Change variables to $y = z/\lambda$ so that the upper limit of the integral is just t, and

$$G(t) = \int_{0}^{t} \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma(k)} dy$$

Therefore the p.d.f. of t is

$$g(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)} \quad ; \ 0 < t < \infty$$

$$= 0 \qquad \text{elsewhere.}$$

This is called a *gamma* distribution. It is the distribution of the total time to exactly k events.

We just quote without proof:

$$E(t) = k/\lambda$$

$$V(t) = k/\lambda^2$$

The exponential is the special case k = 1.

The Normal or Gaussian Distribution

To help motivate this, will study the Central Limit Theorem first. By way of introduction, we have the *Weak Law of Large Numbers.*

If we have n identically distributed, independent random numbers X_i, and if the mean μ and variance σ^2 of the underlying distribution exist, then the r.v.

$$Y_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

converges (in probability) to μ as $n \to \infty$. (The strong law says the same thing, but holds under more general conditions than what we shall show, and also shows a stronger form of convergence than "in probability").

Proof:

Since all the X_i are independent, $V(Y_n) = \frac{1}{n}\sum V(X_i) = \frac{\sigma^2}{n}$.

Therefore, by the Chebychev Inequality, the probability that Y_n differs from μ by more than a pre-specified amount shrinks to zero as n increases:

$$P[|Y_n - \mu| \geqslant \epsilon] \leqslant \frac{1}{\epsilon^2}\frac{\sigma^2}{n} = \frac{1}{\epsilon^2} V(Y_n) \; .$$

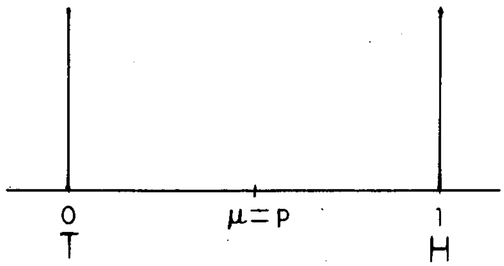
This converges to any arbitrarily small probability for a given ϵ by increasing n. This type of convergence is called convergence in probability. (The strong law has convergence "almost surely," or "almost everywhere").

In fact, there is convergence even if σ^2 is not finite in certain cases, but we won't go into this.

The law of large numbers means that the sample mean is a useful quantity to use in the statistical problem of estimating the true population mean. As n increases, the sample mean converges to the true mean.

Because Chebychev's inequality must hold for any distribution, it is usually very conservative for specific cases. Therefore convergence in this theorem may be more rapid than implied by the proof. This law implies a certain long-run stability in the relative frequency of occurrence of any event in a sequence of trials of the experiment. Look at a rather extreme case: a coin toss. This is a Bernoulli process:

$$p = 1/2 \; (H), q = 1/2 \; (T).$$



All the probability is concentrated at the ends. But the mean of one trial is in the middle.

The expectation value of

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is $\frac{np}{n} = 1/2\ (=p)$, for n tosses, where $X_i = 0$ or 1 for each toss.

By the Chebychev inequality, how many trials do we need in order to have the average Y_n within a specified interval about p with some specified probability? We want the probability P_o that Y_n differs from p by $< \epsilon$, where ϵ is some specified positive number. That is,

$$P[|Y_n - p| < \epsilon] \geqslant P_o \ ,$$

Therefore

$P[|Y_n - p| \geqslant \epsilon] \leqslant 1 - P_o$, which means we must set $1 - P_o = \frac{1}{\epsilon^2} \frac{\sigma^2}{n}$, by Chebychev.

We therefore need the variance:

$\sigma^2 = pq$ for one toss .

By the addition law for variances for independent tosses,

$$V(Y_n) = V\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum V(X_i) = \frac{1}{n^2}\, n \times pq = \frac{pq}{n} \text{ ; and}$$

$$P_o = 1 - \frac{1}{\epsilon^2} \frac{pq}{n} \ .$$

Let $\epsilon = 0.05$ and $P_o = 0.68$:

$$n = \frac{pq}{\epsilon^2} \frac{1}{1 - P_o} = \frac{1/4}{(0.05)^2} \frac{1}{0.32} \cong 310$$

tosses needed to get Y_n within ±0.05 of 1/2 with probability 68%. Check:

$$P[0.45 \leqslant Y_n \leqslant 0.55]$$

$$= P[310 \times 0.45 \leqslant r \leqslant 310 \times 0.55] = P[140 \leqslant r \leqslant 170] \ ,$$

where r is the Binomial parameter = the number of heads;

$$= \sum_{r=140}^{170} \binom{310}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{310-r}$$

$$= \left(\frac{1}{2}\right)^{310} (310)! \sum_{140}^{170} \frac{1}{r!(310 - r)!} \ .$$

Take logs, using Stirling's approximation:

$$\ell n x! \cong -x + x\,\ell n x + \frac{1}{2}\,\ell n(2\pi x)$$

and find P ≅ 0.92. Clearly, even in this rather extreme case, Chebychev's Inequality was overly conservative (0.92 instead of 0.68)

We can also relate this to the standard deviation:

Demand that $\sigma(Y_n) = 0.05$, i.e., 0.05 is one standard deviation (about 68% of the probability if Y_n were Gaussian);

Therefore

$$\sqrt{\frac{pq}{n}} = 0.05 \text{ , hence } n = \frac{pq}{(0.05)^2} = 100$$

This should be more reasonable.

Although it is commonly assumed that ± one standard deviation should cover about two-thirds of the probability, for the Binomial this may be far wrong. For one toss, there is exactly probability 1.0 that Y_1 is within $\pm\sqrt{pq} = \frac{1}{2}$ of $\mu = \frac{1}{2}$. For two tosses, this changes to a probability of $\frac{1}{2}$ that Y_2 is within $\pm 1/(2\sqrt{2})$ of $\frac{1}{2}$. As the number of tosses n increases, the probability that n is within $\pm\sqrt{pq/n}$ of $\frac{1}{2}$ converges fairly quickly to 68.3%, which is appropriate for the Gaussian distribution we are about to consider. It is this sort of convergence to a Gaussian, and the consequent importance of the Gaussian, that is the subject of the:

Central Limit Theorem

This theorem is "central" to statistics. It tells us more than that the average converges to the population mean. It tells us how the average is distributed.

We know that the average becomes arbitrarily tightly distributed about the true mean. Therefore, we *standardize* to a variable that will neither shrink to a δ-function nor expand to an infinite blob:

First, define $Y_n = \frac{1}{n} \sum X_i$,

X_i a r.v. distributed according to almost any p.d.f. with mean μ and finite variance σ^2. Then take

$$Z_n = \frac{Y_n - \mu}{\sigma/\sqrt{n}} \quad ;$$

since $E(Y_n)$ is μ, this has mean 0. Since $V(Y_n) = \sigma^2/n$,

$$V(Z_n) = V\left(\frac{Y_n}{\sigma/\sqrt{n}}\right) + V\left(\frac{\mu}{\sigma/\sqrt{n}}\right)$$

$$= \frac{n}{\sigma^2}\, V(Y_n) = 1 \; .$$

Therefore Z_n has fixed mean and variance independent of n, and we can examine Z_n as $n \to \infty$.

Then the statement of the C.L.T. is

$$\lim_{n \to \infty} P(Z_n \leqslant z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-1/2u^2} du \equiv \mathrm{erf}(z) \ .$$

The function erf(z) defined by this equation is called the *error function.* We will not prove the C.L.T. Almost any of the texts mentioned in the back will supply a proof. The proof requires the use of characteristic functions, which, in the interest of brevity, we do not cover in these notes.

NOTE: the error function is sometimes defined as

$$\phi(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2} dt \ .$$

The relationship of this form to ours is: $\mathrm{erf}(z) = \frac{1}{2} + \frac{1}{2}\,\phi(z/\sqrt{2})$.

The error function is the c.d.f. of the *standard normal,* whose p.d.f. is therefore

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

Since this is symmetric about 0, the mean = 0; we have already derived that the variance = 1. We can, of course, have non-zero mean μ and variance σ^2 not equal to 1:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \equiv N(x\,;\mu\,,\sigma) \ .$$

This is the distribution law obeyed by the average, Y_n, above, as $n \to \infty$. We will use the terms "Gaussian" and "normal" interchangeably to refer to this distribution.

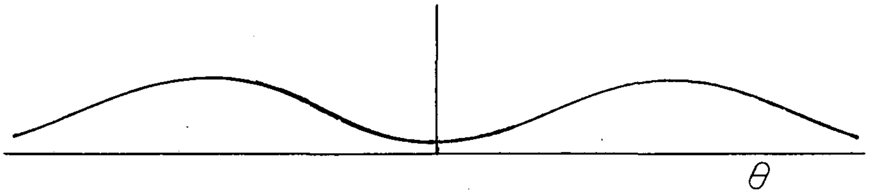
Lots of measurement distributions seem to be approximately normal; e.g., in measuring the length of a table there are contributions to the errors from lots of small effects, such as the behavior of individual muscle fibers, effects due to surface roughness of the table as our ruler slides along it, random vibrations, etc. Any time an error can be regarded as due to the sum over lots of small effects, expect the final measurements to be approximately normal.

Even though each type of effect may be described by a different normal in the C.L.T. limit, the net effect equals the sum of these, and we will see that the sum of a set of normal r.v.'s is itself a normal (reproductive property). Hence, the fact that there is a variety of contributions to the error is not important—what is important is that each type of contribution contributes a small amount a large number of times to *each* measurement.

The case when there exists a few large effects changes this; for example, the penetration of charged particles in matter, undergoing lots of small energy-loss collisions. This yields a fairly tightly distributed range. A few particles undergo large energy-loss deflections, which are more catastrophic collisions. This gives straggling. Early tests (before track detectors) for the nature of rays—α, β, γ—were based on this fact. That is, neutral particles tend to undergo only the large collisions or none at all, giving a wide distribution of ranges. The α and β particles had tightly defined ranges; the x-rays did not.

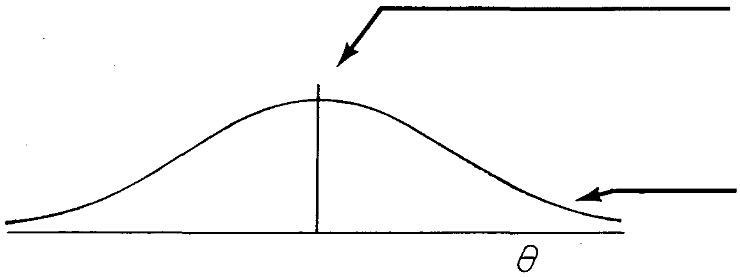
The angular deflection in multiple scattering works this way also. If the net deflection is small, then it is likely due to a very large number of small scatters. If the net deflection is large, then we probably have the same small scatters plus a few large scatters, and we expect deviations from the Gaussian shape, due to the small *number* of large angle scatterings.

What we have then is a lot of scatters in any small range θ to $\theta + d\theta$, θ small. Each of these gives us a Gaussian. The result is the sum of all these which also yields a Gaussian, even though the σ^2 is different for each value of θ [the number of scatters in (θ to $\theta + d\theta$) is Poisson, with parameter depending upon θ according to the laws of Rutherford scattering]. However, for large θ, there are very few scatters, for example, none or 1 or 2. Therefore, the distribution of the end result of large angle scatters only might look like:



(omitting the case of no such scattering)

There are not enough of these per track to approach a Gaussian. Therefore the final net deflection θ might look like:



nearly Gaussian here, where the few large-angle scatters are swamped by the many tracks which have only small-angle scatters.

long tails in the region where it is less probable to get from the many small-angle scatters than from the occasional large-angle scatter.

Caution, or Murphy's Rule of Real Life

This is *often* true, i.e., that we get longer tails than we expect, because of rare contributions of large magnitude to our process. Therefore, simulations of, e.g., detectors, should take account of possible non-Gaussian tails in any random process. If what goes on in the tails of our random processes is important (for example, as a source of background to some other process), we need to be aware that there can be many more occurrences there than Gaussian models would predict. Depending on where we consider the "tails" to lie, this effect can be an order of magnitude or more.

To recapitulate, the *standard* form of the Gaussian or normal is

$$N(x\,;0\,,1) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \;.$$

Then

$$N(x\,;\mu\,,\sigma) = \frac{1}{\sigma}\, N\left[\frac{x-\mu}{\sigma}\,;0\,,1\right]$$

Also,

$$\int_{-\infty}^{z} N(x\,;\mu\,,\sigma)\,dx = \text{erf}\left[\frac{z-\mu}{\sigma}\right] \;,$$

since erf is defined for the standard form. The standard form is found in tables.

Some useful numbers. The probability of a Gaussian r.v. exceeding ±k standard derivations from its true mean is:

1/2	if k is	0.67
.317		1
.046		2
.0027		3
.00006		4

These numbers come from the error function.

Gaussian Approximation to the Poisson

It is often important for what we do that the Poisson can be approximated by the Gaussian in the limit of large numbers. Begin with

$$P_\mu(n) = \frac{e^{-\mu}}{n!}\,\mu^n \;.$$

The mean of n is μ. Assume $\mu >> 1$, and look at n close to μ, i.e., look at $x = n - \mu$; $n >> |x|$. We will also require $x >> 1$, which is usually satisfied if μ is really large.

$$P_\mu(n) = P_\mu(x+\mu) = \frac{e^{-\mu}\,\mu^{x+\mu}}{(x+\mu)!} \;.$$

$$= \frac{e^{-\mu}\,\mu^\mu}{\mu!}\left\{\frac{\mu}{\mu+1}\cdot\frac{\mu}{\mu+2}\;\cdots\;\frac{\mu}{\mu+x}\right\} \quad \text{if} \quad x > 0$$

$$= P_\mu(\mu)\,\{\,''\,\} \;.$$

(We could do this similarly for $x < 0$.) Stirling says $\mu! \cong (2\pi\mu)^{1/2}\,\mu^\mu\,e^{-\mu}$, and so

$$\frac{e^{-\mu}\,\mu^\mu}{\mu!} \cong \frac{1}{\sqrt{2\pi\mu}} \quad \text{if} \quad \mu >> 1 \;.$$

Also

$$\frac{\mu}{\mu+1}\cdot\frac{\mu}{\mu+2}\;\cdots\;\frac{\mu}{\mu+x} = \frac{1}{1+1/\mu}\;\frac{1}{1+2/\mu}\;\cdots\;\frac{1}{1+x/\mu}$$

$$\cong e^{-1/\mu} e^{-2/\mu} \cdots e^{-x/\mu} \ , \ |x| << \mu \ ,$$

and so, finally,

$$P_\mu(x + \mu) \cong \frac{1}{\sqrt{2\pi\mu}} e^{-\left[\frac{1}{\mu} + \frac{2}{\mu} + \cdots + \frac{x}{\mu}\right]} = \frac{1}{\sqrt{2\pi\mu}} e^{-\frac{1}{\mu}\sum\limits_{i=1}^{x} i}$$

$$= \frac{1}{\sqrt{2\pi\mu}} e^{-\frac{1}{\mu}\frac{x}{2}(x+1)} \cong \frac{1}{\sqrt{2\pi\mu}} e^{-x^2/2\mu} \text{ since } x >> 1 \ ,$$

which is the Gaussian, if n is interpreted as continuous,

$$= \frac{1}{\sqrt{2\pi\mu}} e^{-\frac{(n-\mu)^2}{2\mu}} \quad \text{with mean} = \mu \text{ and } \sigma^2 = \mu \text{ also.}$$

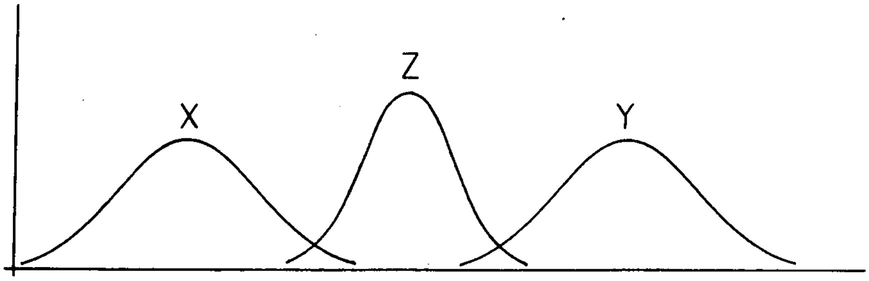
Therefore if we have a bin in a histogram which filled by a Poisson process, then the number of events is approximately Gaussian about its expected number, and the expected variance in that number is the same as the expected number itself.

Reproductive Property of the Gaussian

A reproductive property is to be expected because the Poisson is reproductive. Let X and Y be independent r.v.'s distributed as

$$N(x \, ; \mu_x \, , \sigma_x^2) \text{ and } N(y \, ; \mu_y \, , \sigma_y^2) \ , \text{resp.}$$

Then we can show that Z = X + Y is distributed as $N(z \, ; \mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$ (this implies also that the average of two Gaussian r.v.'s is Gaussian). We know already that $\mu_z = \mu_x + \mu_y$ and $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$ (that's true for any p.d.f., for independent r.v.'s); what we don't know already is that Z is normal.



The proof is tedious, but straight forward. The same result holds for Z = X − Y.

By extension, the sum of any n Gaussian r.v.'s is Gaussian. That is, the C.L.T. holds for any n, in this case.

Higher moments:

$$E[(X - \mu)^n] = 0 \qquad n \text{ odd}$$

$$= (2k - 1)!!\sigma^2k \qquad n = 2k$$

Remember: $(2k - 1)!! \equiv 1.3.5...(2k - 1)$,

e.g., $E[(X - \mu)^4] = 3\sigma^4$
(=> Kurtosis = 0)

Bivariate Normal

Take 2 variables which are themselves normal, but not necessarily independent. Then the distribution of these is

$$f(x_1 , x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \quad \times$$

$$\exp\left\{\frac{-1}{2(1 - \rho^2)}\left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}\right]\right\} \; ; \; \rho^2 < 1 \; .$$

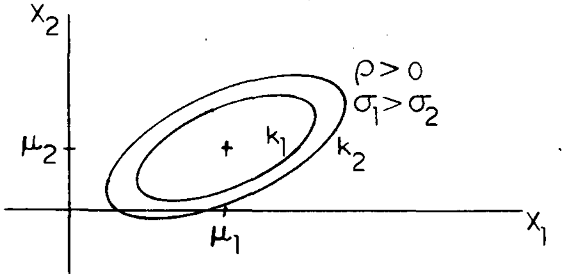
This is the bivariate normal. It is a function of five variables: μ_1, σ_1, μ_2, σ_2, and ρ; it is necessary that σ_1 and σ_2 be > 0 and $|\rho| \leqslant 1$. If $\rho = 1$, the variables are degenerate, being linear combinations of each other.

If we want a contour of constant probability density, i.e., $f(x_1, x_2) =$ constant, then define

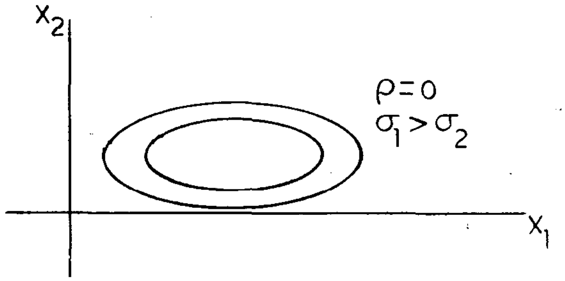
$$G = \frac{1}{(1 - \rho^2)}\left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}\right]$$

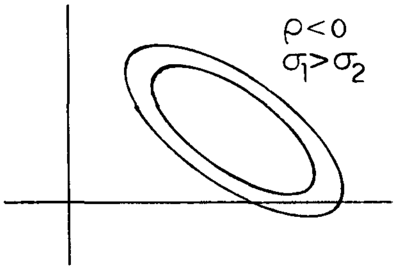
and set G = a constant k.

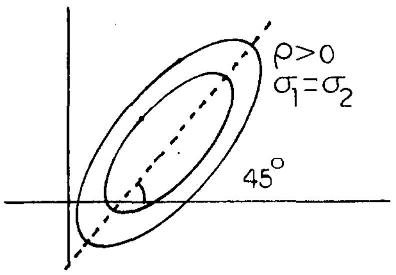
These contours are ellipses.

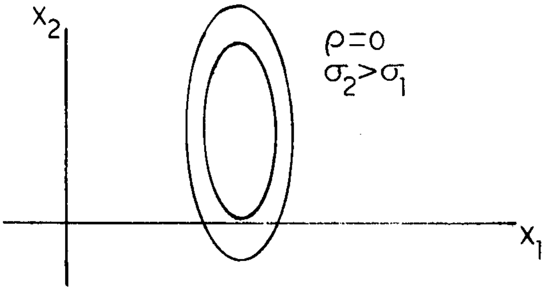


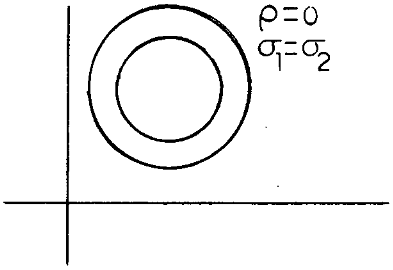
We will show that it is useful to make probability statements of the form $P[(X_1 , X_2) \epsilon$ ellipse]





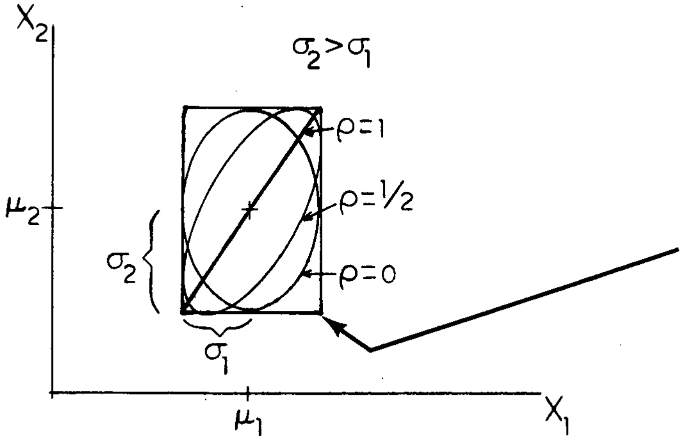






It is clear that ρ plays the role of a correlation coefficient, and it can be shown that it is.

In general,



keep σ_1 and σ_2 constant;
change ρ.

This is the bounding rectangle.
If k = 1, then σ_1 and σ_2
are as shown.

σ_1, σ_2, μ_1, and μ_2 define a bounding rectangle as shown. Then ρ describes the orientation and thin-ness or fat-ness of the ellipse within it. If the two variables are more highly correlated, the ellipse tilts more (one way or the other, depending on the sign of ρ) and also becomes slimmer, so that the probability is more highly concentrated. Note that the angle of the diagonal of the rectangle is not 45° unless $\sigma_1 = \sigma_2$, i.e., this angle depends on the relative errors.

The angle of the major axis of the ellipse is given by θ where

$$\tan 2\theta = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2} \ .$$

Note that $\theta = \pm 45°$ for all $\rho \neq 0$ if $\sigma_1 = \sigma_2$. If, in addition, $\rho = 0$, the ellipse degenerates into a circle whose axes are, of course, undefined. If $\rho = \pm 1$, the major axis is a diagonal of the bounding rectangle [easily demonstrated from $\tan 2\theta = 2 \tan\theta/(1 - \tan^2\theta)$]; otherwise neither axis lies along a diagonal unless $\sigma_1 = \sigma_2$. To invert the above for θ, viz:

$$\theta_1 = \frac{1}{2} \tan^{-1} \left\{ \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2} \right\} ,$$

one must pay careful attention to quadrants. Since one usually defines the arctangent function to lie between $-\pi/2$ and $+\pi/2$, it is customarily said that θ_1 is the angle of the *major* axis if $\sigma_1 > \sigma_2$, otherwise it is the angle of the *minor* axis.

The ellipses at G = k are called covariance ellipses. The meaning of these ellipses is that the probability that a point (x_1 , x_2) sampled from this bivariate normal will lie within the ellipse is given by

$$P[G \leq k] = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \iint_{G \leq k} e^{-\frac{1}{2}G} \, dx_1 dx_2 \ .$$

The ellipses depend upon ρ, as shown in the preceding figure for k = 1. Even though, as $|\rho|$ increases the area obviously decreases, the *probability* within the ellipse remains constant (note: $\frac{1}{\sqrt{1-\rho^2}}$ increases as $|\rho|$ increases) for constant k. The probability just becomes more concentrated. It can be shown in fact that the probability of lying within the ellipse depends only on k and not on μ_1, μ_2, ρ, or even σ_1 or σ_2 (of course, our ellipse will move or change shape if we change those parameters, but the probability of finding a point within it depends only on k). This fact should become clear when we discuss the χ^2 later.

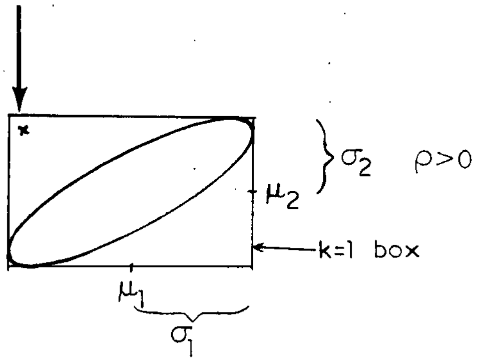
Useful numbers, two-dimensional case:

$P[(X_1 , X_2) \epsilon G = k]$	k
0.39	1
0.63	2
0.78	3
0.86	4
0.92	5
0.95	6

Notice that these probabilities are quite a bit less than the single-variable case we saw before. They are also less than the product of two single-variable cases. This is a k standard deviation ellipse, but it is *not* the same as saying that both X_1 and X_2 are within k standard deviations. The case in which X_1 and X_2 are measured in an experiment, and it is assumed that the random measurement errors are such that the two variables are sampled from a bivariate normal, is a very important application of this model. In this case the covariance ellipse is often called the error ellipse. Error analyses involving

Gaussian variables are discussed in Chapters 10 and 11.

As ρ increases, there is more and more area inside the box which is not included in the ellipse. Therefore, it is a bit misleading to quote as the outcome of a measurement of X_1 and X_2 the one standard deviation limits on each of X_1 and X_2 separately without stating ρ. This is sometimes done, anyway, as a conservative approach (it means one is quoting a larger area of $X_1 - X_2$ space as the "one standard deviation" region, but in a way which depends upon the choice of variables and is therefore not entirely logical), or out of ignorance. We might conclude that points such as

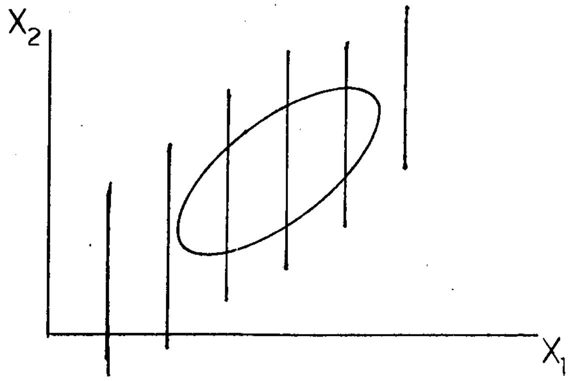


are much more likely than they really are. If ρ is fairly close to ±1, such points could be 5 or more standard deviations from (μ_1, μ_2), yet lie within the quoted "one standard deviation" limits. In any case, where ρ is fairly close to ±1, we should give ρ, or draw covariance ellipses, when quoting errors. If ρ is near zero, the error ellipse will nearly fill the one standard deviation box, and the box is a reasonable approximation to the ellipse. This suggests that we should consider a change of variables to an uncorrelated set. However, that often is impractical, because we want to quote our results in terms of a physically meaningful set, and the correlation is an artifact of the measuring process not necessarily related to the underlying physics.

Conditional Distributions for the Bivariate Normal

Any straight slice through a bivariate normal gives a univariate normal. Therefore, any conditional p.d.f. is normal, with conditional expectation and variance, for example:

$$\left.\begin{aligned} E(X_2|x_1) &= \mu_2 + \rho(\sigma_2/\sigma_1)(x_1 - \mu_1) \\ \\ V(X_2|x_1) &= \sigma^2_{x_2}(1 - \rho^2) \end{aligned}\right\} \text{vertical slices, i.e., the marginal}$$



The profile of the p.d.f. (which is the conditional p.d.f. for X_2) along any slice is the same (in shape), even though the ellipse may be wider or narrower, and even though we may move as far from the central region as we please, because

$$f(x_2|x_1) = \frac{f(x_1\,,\,x_2)}{f_1(x_1)} \quad .$$

Dividing by $f_1(x_1)$ compensates for the fact that we may draw our slice as far from μ_1 as we choose. Of course, unless $\rho = 0$, the *position* of this p.d.f. for x_2 changes with x_1, but its *shape* is unchanged, since the variance does not depend upon the x_1 value of the slice. As always with a Gaussian, the mean suffices to describe the location, once the variance is known. The mean, $E(X_2|x_1)$ lies along a straight line of slope $\rho(\sigma_2/\sigma_1)$, which is neither of the ellipse axes nor the diagonal of the bounding rectangle (for $\rho \neq \pm 1$, of course). It passes through (μ_1, μ_2) and two of the points at which the ellipse is tangent to the bounding rectangle. With $|\rho| < 1$, this line has smaller slope, in absolute value, than that diagonal of the bounding rectangle most nearly aligned with the major axis (which has slope $\pm\ \sigma_2/\sigma_1$); from this, the two points of tangency in question are obvious.

We could take a slice at any angle and still get a univariate normal. The normal we get can be found by making a change of variables such that the slice is along one axis.

Covariance Matrix

Let us now define the *covariance matrix* as

$$V = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} ; \qquad \vec{x} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} \qquad \vec{\mu} = \begin{bmatrix} \mu_2 \\ \mu_1 \end{bmatrix}$$

We shall use the notation for the determinant

$$\det(V) = |V| \ .$$

We assume that V is non-singular, so $|V| \neq 0$. Then it is easy to show, in matrix notation,

$$f(x_1 , x_2) = \frac{1}{2\pi} |V|^{-1/2} \exp\left[- \frac{1}{2} (\vec{x} - \vec{\mu})^T V^{-1}(\vec{x} - \vec{\mu}) \right] ,$$

which displays the formal similarity to the univariate normal. The notation "T" represents "transpose."

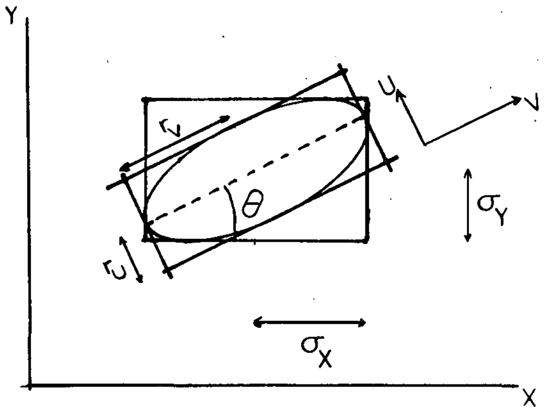
Note that V non-singular requires $\rho \neq \pm 1$. That is, the two variables must not be linear combinations of each other. Otherwise we should make a change of variables and take the marginal distribution, i.e., collapse down to a one-dimensional space, for most purposes.

We may also refer to this as the variance matrix.

Theorems

i) If X and Y are bivariate normal r.v.'s, then it is always possible to find a change of variables to U and V, the new variables being a linear combination of the old, s.t. the new variables are uncorrelated.

To do this, we just diagonalize V, using results from matrix theory. This produces a rotation through an angle θ, without a scale change.



The standard errors will change. The semidiameters will become the standard errors, given by (the square roots of)

$$r_v^2 = \frac{\sigma_x^2\sigma_y^2(1-\rho^2)}{\sigma_y^2\cos^2\theta - \rho\sigma_x\sigma_y\sin 2\theta + \sigma_x^2\sin^2\theta}$$

$$r_u^2 = \frac{\sigma_x^2\sigma_y^2(1-\rho^2)}{\sigma_y^2\sin^2\theta - \rho\sigma_x\sigma_y\sin 2\theta + \sigma_x^2\cos^2\theta},$$

for this choice of U and V axes (of course, ρ refers to the X-Y correlation).

If $\rho > 0$ and therefore $\sin 2\theta > 0$ (since the ellipse tilts up to the right in that case, and $0 < \theta < \pi/2$), then r_v^2 increases with increasing ρ and r_u^2 decreases with increasing ρ. Pictorially, the covariance ellipse becomes longer and thinner. Recall from earlier that θ may be computed from

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$$

(provided we are careful about quadrants).

Note that, if we simply make a rotation, the areas don't change. Therefore, P[(U,V) inside ellipse] = P[(X,Y) inside ellipse]

ii) In particular, we can always find a linear transformation

$\vec{U} = C\vec{X} + \vec{D}$, C a matrix, s.t. the components of $\vec{U}$ are *standardized* independent normals (mean 0, variance 1, correlation 0). The value of G (defined earlier) does not change at the transformed point corresponding to the old point. This transformation takes an ellipse into a circle. In general, this requires a rotation until $\rho = 0$, a translation until $\mu_u = \mu_v = 0$, and a scale change to make the variances unity.

iii) If $\vec{X}$ is bivariate normal with diagonal variance matrix, then the components of $\vec{X}$ are independent. That is, that the correlation is zero is necessary and sufficient for the components of $\vec{X}$ to be independent. This is *not* true in general for other distributions, as we saw.

Multivariate Normal

Let $(X_1,...,X_n)$ be a vector of normal r.v.'s, and let σ_i^2 be their variances, μ_i their means, ρ_{ij} their pair-wise correlations, and define

$$V = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \rho_{n1}\sigma_n\sigma_1 & & \cdots & \sigma_n^2 \end{bmatrix} ; \ |V| \neq 0 \ .$$

(Note, $\rho_{ij} = \rho_{ji}$, therefore V is symmetric). Then

$$f(x_1 , \cdots , x_n) = \frac{1}{(2\pi)^{n/2}} |V|^{-1/2} \exp\left[- \frac{1}{2} (\vec{x} - \vec{\mu})^T V^{-1}(\vec{x} - \vec{\mu}) \right] .$$

is their distribution. This is the *multivariate normal* distribution.

The above three theorems apply here, as well. This assumes V is non-singular, i.e., none of the variables X_i can be expressed in terms of linear combinations of the others. If V is singular, it is often best to make a transformation down to a linearly independent set of variables.

In general, V is positive-definite, so that any quadratic form $y = X^T V X$, X a column vector of coordinates, is positive, and any contour y = constant is an ellipse.

This form for the multi-variate normal suggests a useful generalization of a single-variable problem. Much of the statistics we shall do depends upon the Gaussian. It will often be true that we can generalize from a single-variable problem if we replace

X by $\vec{X}$

μ by $\vec{\mu}$

σ^2 by V

σ^{-2} by V^{-1} (V non−singular)

σ by $|V|^{1/2}$

$\frac{1}{\sqrt{2\pi}}$ by $\frac{1}{(2\pi)^{n/2}}$

This is a rule of thumb, not a law.

The χ^2 Distribution

Let $X_1 , \dots , X_n$ be independent normal r.v.'s (hence, uncorrelated), each with mean μ_i and variance σ_i. Then

$$f(\vec{x}_i \, ; \vec{\mu}_i \, , \vec{\sigma}_i) \, d\vec{x}_i = \prod_{i=1}^{n} \frac{\exp\left[- \frac{1}{2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right] dx_i}{\sqrt{2\pi\sigma_i^2}}$$

$$= \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right] \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \, dx_i \ .$$

We define the variable called chi-squared as

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 .$$

The χ^2 has a parameter, n. "n" is called the number of degrees of freedom (d.o.f.) of the χ^2, since each Gaussian variable x_i is free to vary independently according to the dictates of its own distribution. The χ^2 is a single variable, not the square of a variable; one does not usually refer to a variable "χ" = $\sqrt{\chi^2}$, but it can be done for certain purposes.

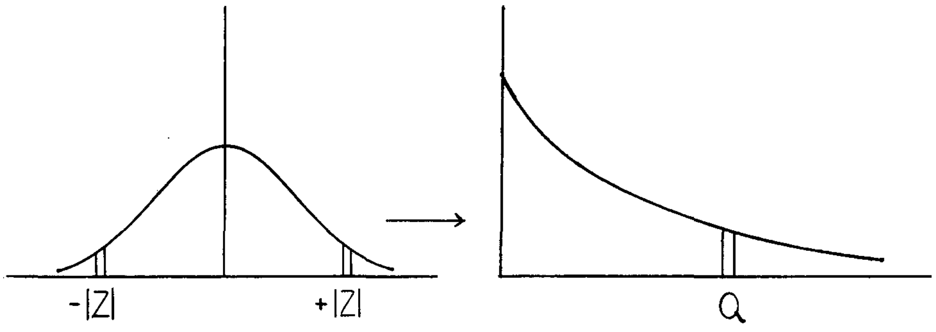
Examples: For n = 1, let $z = \frac{x - \mu}{\sigma}$. Then

$$f(z)dz = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \quad \text{(our normal).}$$

Let $Q = Z^2$ ($\equiv \chi^2$)

$$dq = 2zdz \ , \ \text{therefore } dz = \frac{dq}{2z} = \frac{dq}{\pm 2\sqrt{q}}$$

This is not a 1-1 transformation, because the ±Z region goes into the +Q region only



The probability that Q is between Q and Q + dQ = the probability that Z is between Z + dZ around Z = $\sqrt{Q}$ *and* between Z and Z + dZ around Z = $-\sqrt{Q}$. Therefore

$$f(q)\,dq = \frac{1}{\sqrt{2\pi}} \left\{ e^{-\frac{1}{2}q} \left(\frac{dq}{2\sqrt{q}} + \frac{dq}{2\sqrt{q}} \right) \right\}$$

(the intervals dq or dz are always taken to be positive only)

$$= \frac{1}{\sqrt{2\pi q}} e^{-\frac{1}{2}q} dq \ .$$

$$Q = \chi^2 \Rightarrow f(\chi^2) = \frac{1}{\sqrt{2\pi\chi^2}} e^{-\frac{1}{2}\chi^2} \ .$$

This distribution is the χ^2 with one degree of freedom.

For $n = 3$, $z_i = \frac{x_i - \mu_i}{\sigma_i}$, standardized normal variables. $\chi^2 = Z_1^2 + Z_2^2 + Z_3^2 = R^2$. It is convenient to think of a 3-dimensional volume whose Cartesian axes are Z_1, Z_2, Z_3. Then R^2 is the radius of a sphere. If is left as an exercise to show that

$$f(R)dR = \frac{2}{\sqrt{2\pi}} R^2 e^{-R^2/2} dR \ .$$

Change variables to $\chi^2 = R^2$. Again, χ^2 is a random variable, whereas R^2 is the square of a random variable. That is,

$d\chi^2 = 2RdR$, and

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

This is the χ^2 with three degrees of freedom.

In general,

$$f(\chi^2) = \frac{(\chi^2)^{(n/2-1)} e^{-\chi^2/2}}{\Gamma(\frac{n}{2})\, 2^{n/2}} \quad \text{for n d.o.f.}$$

$\Gamma(\frac{n}{2})$ is the Gamma function:

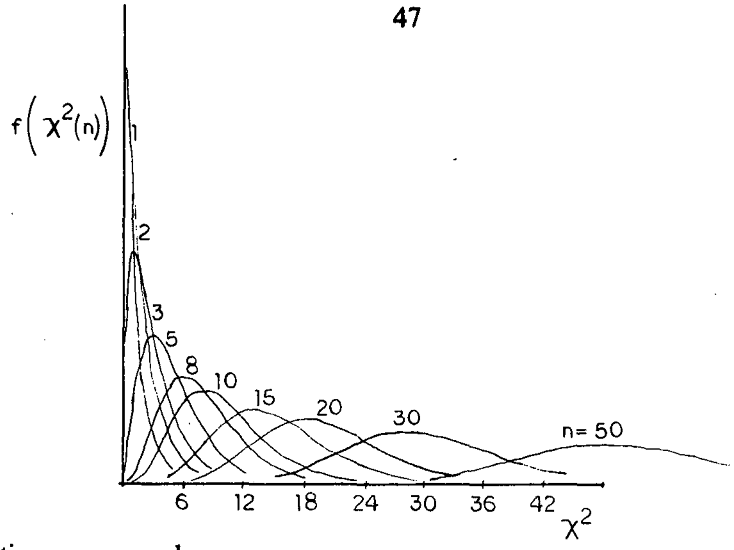
$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

$$\Gamma(n + 1) = n\Gamma(n)$$

$$\Gamma(\tfrac{n}{2}) = (\tfrac{n}{2} - 1)! \ ; \text{n even}$$

$$= \frac{(n-2)!!}{2^{(n-1)/2}} \sqrt{\pi} \ ; \text{n odd}, > 1 \ .$$

Recall, $(n - 2)!! = (n - 2) \times (n - 4) \times ... \times 5 \times 3 \times 1$.



Using Γ functions, we can show

$E[\chi^2(n)] = n$

$\sigma^2_{\chi^2(n)} = 2n$

Mode (peak) = n − 2 (at 0 for n = 1 or 2).

If we have n Gaussian variables X_i which are not independent, (i.e., they are correlated), we can do a change of variables s.t. the variance matrix is diagonalized, without changing the value of χ^2. That is, we remain on the same covariance ellipse $G \equiv \chi^2 = k$. G is defined in the above discussion on the Gaussian. Then the new variables are independent, and therefore we see that $(\vec{x} - \vec{\mu})^T V^{-1}(\vec{x} - \vec{\mu})$ is still a $\chi^2(n)$ even in the case of non-independent variables, unless V is singular. If V is singular, there exists a linear relationship among some of the variables. This is very important in the case of constrained fitting, where the constraints may provide just such a relationship. In this case, if there exist $m < n$ independent variables, where m is the dimensionality of the largest such set, then

$$\sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \quad \text{is } \chi^2(m) \text{ (note: the sum goes to "n")} .$$

There are other derivations of the χ^2, but we see it most often as a sum of variables of this form.

Reproductive Property

Let χ_i^2 be a set of variables, each of which is $\chi^2(n_i)$. Then $\sum\chi_i^2$ is $\chi^2(\sum n_i)$. The proof is straightforward.

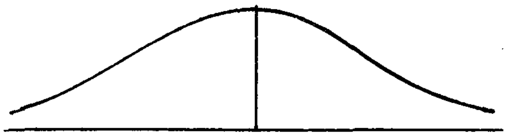
Sometimes one sees χ^2/n quoted as a figure of merit in a test of a hypothesis (e.g., least-squares fitting, to be discussed in Chapter 11). That is, we wish to convey a feeling for the probability of a certain χ^2 value we have observed, in a manner which removes some of the dependence upon n. This is flawed, because values of $\chi^2/n \cong$ a few are of high probability at low n, whereas at high n, a value of, say, 1.5 may be highly unlikely. For example, the probability that $\chi^2/n > 2$ is:

≃ 15% for n = 1
9% for n = 4
3% for n = 10
1/2% for n = 20
$< 10^{-2}$% for n = 50.

One more distribution:

The Cauchy (the Breit-Wigner)

$$C(x; \mu, \alpha) = \frac{1}{\pi\alpha} \frac{1}{1 + (x - \mu)^2/\alpha^2} ; \quad -\infty < x < \infty$$



like a long-tailed Gaussian

$$E(x) = \int_{-\infty}^{\infty} x\, C(x;\mu,\alpha)dx = \frac{1}{\pi\alpha} \int_{-\infty}^{\infty} \frac{x}{1 + (x - \mu)^2/\alpha^2} dx$$

$$= \frac{1}{\pi\alpha} \int_{-\infty}^{\infty} \frac{(x - \mu) + \mu}{1 + (x - \mu)^2/\alpha^2} dx$$

$$= \frac{1}{\pi\alpha} \frac{1}{2} \ell n \left[1 + \frac{(x - \mu)^2}{\alpha^2} \right] \Bigg|_{-\infty}^{\infty} + \mu$$

$$= \mu + \infty - \infty,$$

which is indeterminate. Higher moments likewise do not exist. We can *define* a mean as

$$E(x) = \lim_{L\to\infty} \int_{-L}^{L} x\, C(x;\mu,\alpha)dx$$

$$= \mu + \frac{1}{2} \lim_{L\to\infty} \frac{1}{\pi\alpha} \ell n \left[1 + \frac{(x - \mu)^2}{\alpha^2} \right] \Bigg|_{-L}^{L}$$

$$= \mu .$$

This is the bad boy of all the distributions in common use. No trick will enable us to define a finite variance.

The Cauchy has a strange reproductive property.

Let X_i be a set of independent r.v.'s distributed as $C(x_i,\mu,\alpha)$. Then $\overline{x} = \frac{1}{n} \sum x_i$ is also $C(\overline{x};\mu,\alpha)$ i.e., the *same Cauchy* as each one separately. Therefore the mean of n Cauchy variables does not converge to anything — because of the long tails. It violates the C.L.T.! The reason is, it has no finite variance.

An example of the Cauchy:

$$C(M;M_o,\frac{1}{2}\Gamma) = \frac{\frac{1}{2}\Gamma}{\pi} \frac{1}{(M - M_o)^2 + (\frac{1}{2}\Gamma)^2} ,$$

the familiar Breit-Wigner. In physical problems the long tails are truncated by energy conservation or something and the resultant distribution is well-behaved in the sense that all moments exist. However, the mean and variance may be sensitive to the exact manner in which the tails are cut off. Since the tails are usually the least interesting and least well-measured portions of the distribution, one frequently uses the peak location and the "full width at half maximum" (FWHM) for measures of location and width. The peak occurs for $x = \mu$, and the p.d.f. is half its peak height at $x - \mu = \pm\alpha$ (yielding FWHM $= 2\alpha$), so that these are rather satisfying measures of location and width. For a Gaussian, the peak is of course likewise at $x = \mu$; the curve falls to 1/2 peak height at $x - \mu = \pm 1.1774\ \sigma$ so that the FWHM is $2.355\ \sigma$.

This is the end of the discussion on distributions. There are other distributions which are useful in certain cases—we will discuss them at the appropriate time. (They are less familiar.)

Chapter 4. Monte Carlo

Let us digress for a moment into *Monte Carlo.* Many of the results of this section may be found in Rubenstein (1981), to which we are indebted for most of the examples. Rubenstein should be consulted for further details and additional topics.

Any statistical problem for which probabilities can be estimated (or, in some cases, at least guessed at) can be simulated on a computer. In principal, this includes problems of arbitrary complexity, in which the probabilities at one stage depend upon exactly what happened at earlier stages. The creation of high energy physics reactions and subsequent propagation of the particles through a detector with a complex configuration is an example. One can then study the response of the detector as a function of physical quantities of interest at the moment of creation. Another example involves the physics of the creation of these events. One may have a probabilistic model for creation in terms of certain variables and need to look at its predictions in terms of a large number of other variables. One may need to do this after removing certain portions of the data necessitated by the experimental conditions (e.g., portions with large backgrounds), and after incorporating the characteristics of the detector.

The correct p.d.f.'s for such processes are typically completely unobtainable by direct calculation. Monte Carlo programs for extremely complex cases may sometimes take time to set up, and often still involve assumptions untestable except by the final result. However, it is often true that even moderately complex cases can be set up quickly, perhaps using already-available program sub-units. Even in simple cases a one-page Monte Carlo can provide valuable insight into a problem, for example revealing sources of errors in attempts at direct calculation.

These processes must be broken down into a number of separate random processes. Each subprocess is simulated with the use of random numbers "generated" according to a certain distribution. The basis of this in most cases is a "pseudo-random" number generator, a function which returns a number (a "uniform deviate") claimed to be sampled from a *uniform* distribution on the range 0 to 1.

Definition: *pseudo-random:* The numbers satisfy various statistical tests for random numbers uniform between 0 and 1, but they are generated by a reproducible process. The chain of numbers is initiated by some starting number X_0, which may be kept the same from run to run or changed (e.g., to the time of day in milliseconds) to provide a "random" starting point. The starting number X_0 is not used as one of the random numbers, but is used in the calculation of the first random number. That first number is then usually used in the calculation of the second, and so on.

All pseudo-random number generators will fail a "truly randomness" test somewhere—i.e., they all have problems. For the most part, these problems are tolerable, but it's worth checking one's results closely. It is not uncommon to find errors due to lack of randomness. In some cases, you may need to switch to a slower but better generator. Truly-random number generators can be constructed. They are based upon naturally-occurring random processes, e.g., radioactive decay. These are very difficult to make, and can be wrong. Numerous psychics have built such devices and claimed extra-ordinary powers based upon their ability to interrupt them more frequently at certain numbers than statistics would predict. In spite of this testimonial, correct devices have been built and used for certain applications.

Our problem is to transform this uniform deviate to a number sampled from some other distribution on some other range. There are three basic methods employed:

Inverse Transform Method
Composition Method
Acceptance-Rejection Method

Single-Variable Case

Inverse Transform Method

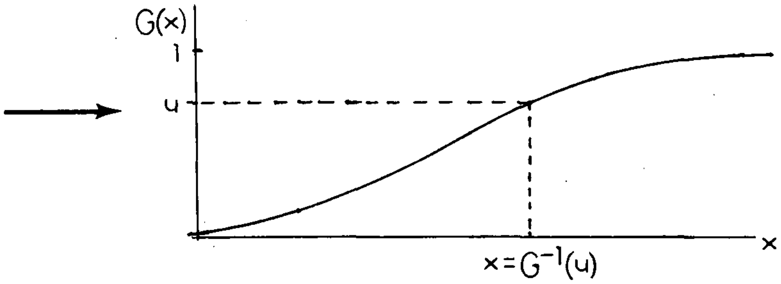
In the change-of-variables discussion, we considered a p.d.f. as a function of one variable u and re-wrote the problem in terms of a p.d.f. for a functionally-related variable x. There we saw that to go from u → x, where $\int f(u)du = 1$, g(x) is (almost) any p.d.f., we wrote $g(x) = f[w(x)] \left|\frac{dw}{dx}\right|$, where if x = v(u), then u = w(x).

For the *continuous* case, if G(x) is the c.d.f. of the desired variable x, then, for u a random deviate uniform on [0,1] returned by our generator, let us choose

$$x = G^{-1}(u) \ .$$

Therefore u = G(x) and G is our inverse function w. This is the general solution if G^{-1} is a one-to-one transformation, because f(u) = 1. Then g(x) will be the uniform p.d.f. times the Jacobean, that is $1 \times \left|\frac{dw}{dx}\right| = \left|\frac{dG}{dx}\right|$ (at least, where everything is > 0), as desired. So, if we have a uniform random deviate u, choose $x = G^{-1}(u)$ to get a random deviate x distributed according to g(x). Then $P(X \leqslant x) = P[G^{-1}(U) \leqslant G^{-1}(u)] = P[U \leqslant u] = u$ (since u is uniform on [0,1]) = G(x).

We choose u uniformly along here. Then,

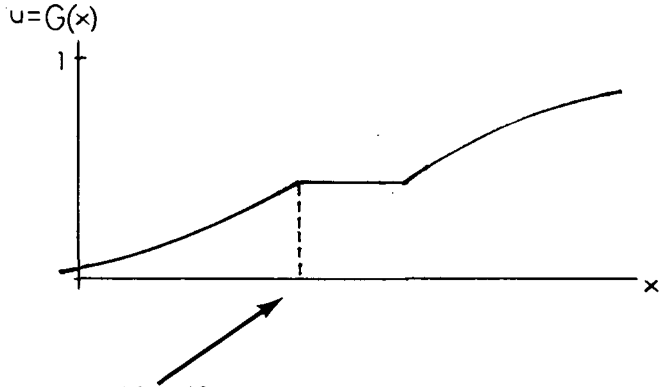


that gives x along here, according to g(x).

If the slope of G is large, then a fairly large range in u on the vertical axis will choose values from a fairly small region of x, and vice versa if the slope is small. This is just what we want because if the slope is large, a small increase in x increases the c.d.f. by a lot, which means there is a lot of probability in that range.

If G is not one-to-one, we *define*

$$x = G^{-1}(u) = \text{smallest}\,\{x : G(x) \geqslant u\}\,.$$



e.g., the case with plateaus

We choose this x if u happens
to put us on the plateau

The Inverse Transform Method is the most straightforward approach. It is usually, but not always, the fastest approach in terms of computer time.

Example: Exponential.

$$f(x) = \frac{1}{\beta}\, e^{-x/\beta}; \qquad 0 \leqslant x < \infty; \beta > 0$$

$$F(x) = 1 - e^{-x/\beta}\ .$$

F(x) must be chosen with equal probability in each equal interval dF. We will set F(x) = u, a uniformly distributed r.v., and solve for x. Then $u = 1 - e^{-x/\beta}$ and $\ell n\,(1 - u) = -x/\beta$. Therefore, $x = -\beta\, \ell n\,(1 - u)$ will be exponential.

Note: $1 - u$ is just another uniform deviate on the same range. Therefore, we can save a little effort by using $x = -\beta\, \ell n\, u$. This yields the same distribution for x.

Example: Cauchy.

$$C(x;\mu,\alpha) = \frac{1}{\pi\alpha}\ \frac{1}{1 + (x - \mu)^2/\alpha^2}\ ;$$

The c.d.f. of the Cauchy is $F(x) = \frac{1}{2} + \tan^{-1}\left(\frac{x - \mu}{\alpha}\right]$. Therefore, set u = F(x) and solve for x, which means we generate

$$x = F^{-1}(u) = \mu + \alpha\tan\,[\pi(u - \frac{1}{2})] = \mu - \frac{\alpha}{\tan(\pi u)}\ ,$$

since $\tan\,(\frac{\pi}{2} \pm \alpha) = \mp \cot \alpha$.

For *discrete* distributions.

Let $P(X = x_k) = P_k$, k an index = 0, 1,... Then the c.d.f. of this discrete distribution is

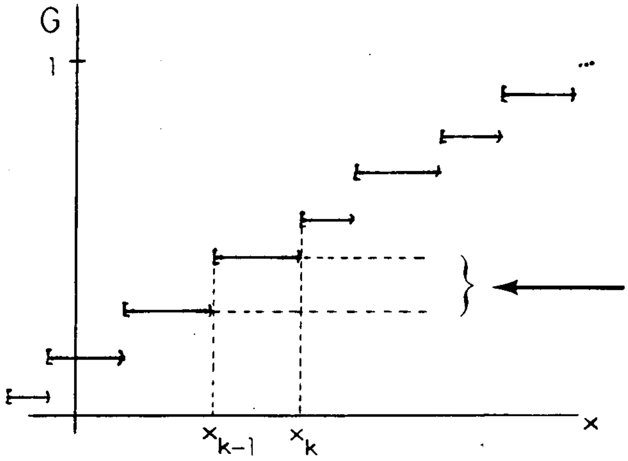
$$G_k = P(X \leqslant x_k) = \sum_{i=0}^{k} P_i \leqslant 1 .$$

Note that this summation includes i = k.

Then

$$P(G_{k-1} < U \leqslant G_k) = \int_{G_{k-1}}^{G_k} du = G_k - G_{k-1} = P_k .$$

(Note: $G_{-1} = 0$)



P_k is precisely this jump, the probability that x is less than x_k but greater than or equal to x_{k-1}.

Take $x = \{x_k : G_{k-1} < u \leqslant G_k\}$

It means we have to

1) generate u (think of it as being along the vertical axis, as in the continuous case);

2) compare u with G_k until we find the correct interval ($G_{k-1} < u \leqslant G_k$), then take the appropriate x_k (it's the upper one).

(There's always a $G_k \geqslant$ any u you take, since $u \leqslant 1$ and $G_k \to 1$ for some k.)

Step 2) may require a lot of comparisons. You can often save computer time by starting, not at k = 0, but at some value in the mid-range, e.g., the mode (the most popular value), and then working up or down according to u and G_k. We won't cover any examples here, because they are time-consuming, and not especially enlightening. Most of the important distributions have tricks which are unique to their own cases.

This concludes our discussion of the inverse transform method. For cases which are, or can be represented as, sums of p.d.f.'s, we can use the

Composition Method

Here, we decompose the desired distribution, which may be difficult or time-consuming, into two or more others which are relatively easy to do. For example:

$$f(x) = \frac{5}{12}[1 + (x - 1)^4]; \quad 0 \leqslant x \leqslant 2 \ .$$

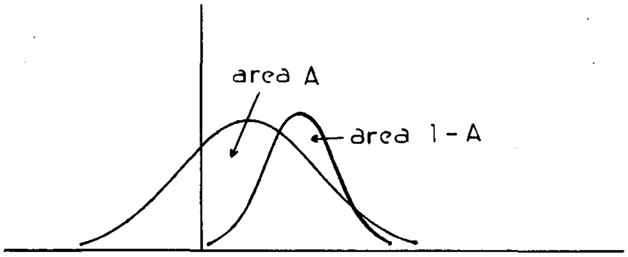
Use the composition method to generate this distribution. We will de-compose this into a sum of 2 other p.d.f.'s.

Write $f_a(x) = \frac{1}{2}$; $f_b(x) = \frac{5}{2}(x - 1)^4$; $\quad 0 \leqslant x \leqslant 2$

Each of these is a p.d.f. (including normalization).

Then $f(x) = \frac{5}{6} f_a(x) + \frac{1}{6} f_b(x)$. This is still a p.d.f. The coefficients must sum to 1.

This procedure is like adding two curves, e.g., Gaussians:



The net result must be the sum of the two. Since this net result must be normalized to unit area, the areas must sum to 1.0. We can choose our random numbers separately out of each one. For each random number, we must choose it from Curve (1) with probability A and from Curve (2) with probability 1 − A.

Now, back to our example, we generate *two* uniform random deviates, u_1, u_2 uniform on the interval [0,1]:

u_1 *chooses* between f_a or f_b;

u_2 *inverts* the chosen f_a or f_b, using the inverse transform method.

Thus,

if $u_1 < 5/6$, $x = 2\,u_2$ (inverting f_a)

if $u_1 \geqslant 5/6$, $x = 1 + \sqrt[5]{2\,u_2 - 1}$ (inverting f_b)

Note: Since $F = F_a + F_b$, we could have done it in one step, but the two-step process means we only have to evaluate the 5th root 1/6th of the time. If we had to do this millions of times, it might save significant amounts of computer time, even though we must generate two random numbers to get one out.

Example:

$$f(x) = n\int_1^\infty y^{-n}e^{-xy}dy\ ; \quad n \geqslant 1\ .$$

The variable y is a dummy, which we shall make use of, as we shall see. Inverting this directly is very difficult or impossible. We will write this f as a *probability mixture* of selected density functions, each of which we can do. The key step is to write $f(x) = \int_0^1 g\,(x|y)\ dF\ (y)$. This is a marginal p.d.f. Referring back to the discussion on marginal p.d.f.'s (Chap. 2), we have

$$f(x,y) = f(x|y)\ f_2(y)\ \ .$$

$$\text{Therefore}\ \ f_1(x) = \int f(x|y)\ f_2(y)\ dy$$

$$= \int f(x|y)\ dF\ (y)\ ,$$

which is a marginal p.d.f., displaying the desired form. We have changed (dummy) variables from y to F, using $dF/dy = f_2(y)$ as the Jacobean. F is the c.d.f. for y.

Whereas in the previous example we had two curves and we chose one or the other, here we have a continuously variable family of curves, g. Instead of an area A which tells us the probability with which to choose a given curve, we have a continuous probability function dF which tells us the probability of choosing the parameter y in any certain interval. The parameter y, which we will choose according to this prescription, selects the exact curve g. The integral over dF, then, replaces the sum over a countable number of curves.

We choose a value for y first. This tells us what g to use. Then from g we choose x.

Here, we can take $dF(y) = \frac{ndy}{y^{n+1}}$; $1 < y < \infty$, and $g(x|y) = ye^{-yx}$. We put a factor of y into $g(x|y)$ so we would have a nice exponential p.d.f. for x (once y is known). Then $F(y) = y^{-n}$. When changing variables from y to F, we use the absolute value of the Jacobean so that dF is positive. If we wanted to avoid this complication, we could use $F'(y) = 1 - y^{-n}$, so that F′ goes from 0 to 1 as y goes from 1 to infinity. This gives the same dF and the same random variables. When going from an integral over y to an integral over F, we must change the direction of integration because $y = 1 \Rightarrow F = 1$; $y = \infty \Rightarrow F = 0$; i.e., positive increments in y result in negative increments in F — but dF must represent a probability increment, which must therefore be positive. Here is what happens to the minus sign:

$$n\int_1^\infty y^{-n}e^{-xy}dy = \int_1^0 ye^{-xy}d(y^{-n}) = \int_0^1 ye^{-xy}dF\ .$$

The minus sign is swallowed by changing the direction of integration, i.e., by integrating in the $-y$ direction.

Choose u_1 uniform on [0,1] and set it equal to $F = y^{-n}$; then, since F is the c.d.f. of the marginal for y, we can set $y = u_1^{-1/n}$ by the inverse transform method (in terms of $F' = 1 - F$, the same selection will do, since both u_1 and $1 - u_1$ are equally valid uniform r.v.'s on [0,1]). We have now selected our dummy variable y which chooses the particular exponential p.d.f. from which we are going to select x. We just want the next random number from this family of curves. Therefore, having selected y and therefore a particular curve, select x:

$g(x|y) = ye^{-yx}$ exponential with decay constant y, therefore

$x = -1/y \ell n\, u_2$ will do the trick.

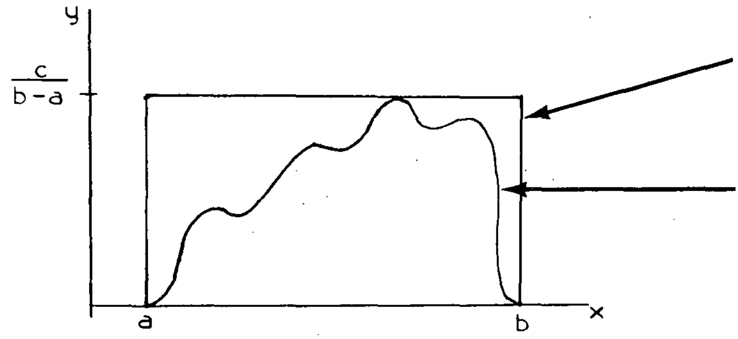
Discussion

In the general case, the integral over dF is a sum over all the possible functions g. Each g is weighted by its probability. If we pick enough random numbers, we will have an accurate representation of the integral. Even if we pick only one random number, we will have a representation of the integral—how accurate will depend on how much g varies over the region of y for which the probability of selection is large. This in turn will depend upon x, so some regions of x will be more accurately represented by a given number of random numbers than other regions. We will see later (Chapter 7) that y corresponds to a *sufficient statistic,* because picking y is sufficient to completely specify the member of the family; then we pick x from that member of the family.

This has fundamental parallels with the general problem in statistics (parametric statistics, really). The situation is reversed: we start with all the x's known. We want to determine the underlying p.d.f. g, which is evaluated at fixed y because y is a constant of nature. To determine g we determine y, assuming g belongs to a certain family of p.d.f.'s with y as a parameter. The difficulty occurs because the x's have all the randomness coming from nature's selections of u_2, a completely uninteresting variable. We must find ways of seeing through this disorderliness to find a range of values of y which have reasonable probability of covering the true value of the constant of nature.

Acceptance-Rejection Method (Von Neumann)

This is the most important method for problems which are complex, that is, not easily represented by a single p.d.f. which one can write down. For example, studies of detector efficiency, of complex theoretical shapes, etc. often fall in this category. It applies to cases with continuous variables.



a constant × a uniform p.d.f.

our p.d.f., f(x); or, a function not normalized, but proportional to f(x). We desire a r.v. X distributed according to this p.d.f.

Choose a p.d.f. uniform on the interval (a,b), the limits of our p.d.f. If a and b are not known, we must bracket them by an interval which is large enough. Find a constant C such that C times this uniform p.d.f. is everywhere $\geqslant$ f(x). Again, if the maximum of f(x) is not explicitly known, we must choose a constant large enough. Then, first, generate an x uniformly on (a,b). Then generate a y uniformly on (0,C/(b − a)), as shown. So far, our (x,y) points will populate the box uniformly. Evaluate f(x). If y $\leqslant$ f(x), accept x as the next value of the random number. If y > f(x), reject x and try again.

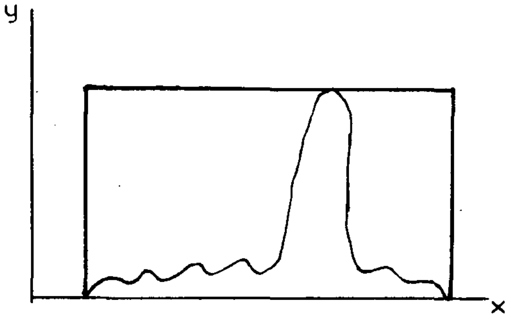
In this way we carve the desired p.d.f. out of the uniform p.d.f. by accepting only (x,y) pairs which lie under the curve. We accept a·fraction

$$\text{efficiency} = \frac{\text{area under p.d.f. or f(x)}}{\text{area under bounds}}$$

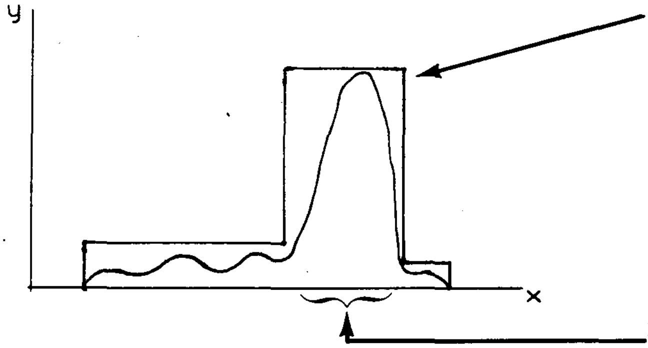
of our trials.

Importance Sampling

This is the complete method. However, sometimes the efficiency is very low. We might not know the maximum of f(x) and therefore have to take an inefficient bound. Or, f(x) may have sharp peaks such that the rectangle leaves out a lot of area:



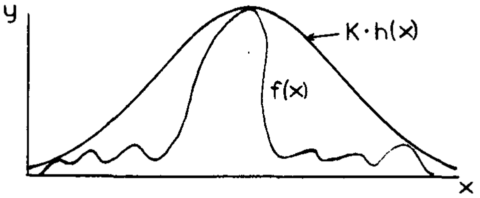
Then we do better by choosing some other bounding p.d.f. (other than uniform). For example: piecewise uniform:



This is proportional to the p.d.f. we use to choose the initial x.

We generate a higher density of x's here.

or a curve, if one can be found which we can invert easily:



K > 1 if f is normalized.

h(x) is a normalized p.d.f.

Now we choose x according to the p.d.f. represented by this curve, h(x). Once chosen, we then choose y *uniformly* as before except on the interval (0, K · h(x)), and test as before. If y ⩽ f(x), deliver x. The method of sampling x in a variable way, responding approximately to the variations in f(x), is called *importance sampling.* It is left as an exercise to prove Von Neumann's

Theorem: Let X be a r.v. distributed with a p.d.f. f(x), x defined on some range. Represent f as

$$f(x) = Cg(x)h(x)\,,$$

where $C \geqslant 1$ is some constant, h(x) is a p.d.f. (chosen such that we can easily generate random numbers from it), and g(x) is some function, $0 < g(x) \leqslant 1$. That is, $Ch(x) \geqslant f(x)$, and g is the fudge factor to correct Ch(x) into f(x). Let U and V be independent r.v.'s distributed as uniform on [0,1] and as h(v), resp. Generate a (u,v) pair. Then the distribution of v *under the condition* that $u \leqslant g(v)$, is identical to the distribution of x:

$$h(v|\, u \leqslant g(v)\,) = f(v).$$

If $u \leqslant g(v)$, then v is returned as the "next value of x"; otherwise we try again with a new (u,v) pair. The proof requires Bayes' Theorem (q.v., Chapter 2):

$$h(v|u \leqslant g(v)) = \frac{P(u \leqslant g(v)|v = x)h(x)}{P(U \leqslant g(V))}\ .$$

The denominator is the integral of the numerator over all possible values of V = x, which we must show is given by

$$P(U \leqslant g(V)) = 1/C\,.$$

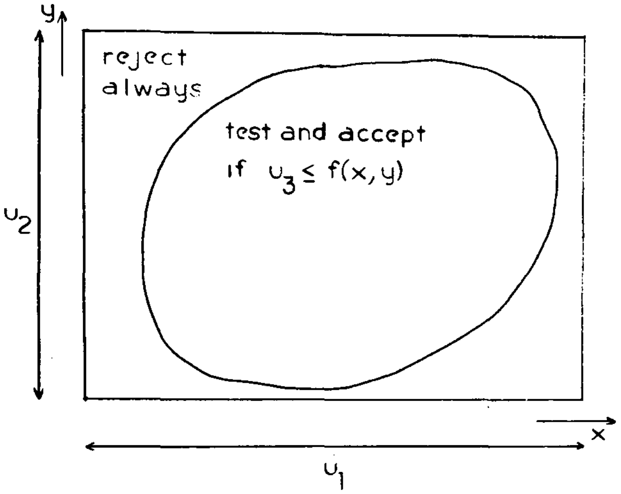
This is the probability of acceptance of v, taken over the set of all possible v's. Therefore the efficiency is p = 1/C, which reflects the fact that we want to choose h(x) as close to f(x) as possible, so that C is as close to 1 as possible. If n is the number of trials before a successful one (i.e., we are successful on trial n + 1), the p.d.f. of n is

$$P(n) = p(1 - p)^n\ ;\ n = 0,\ 1,\ \cdots$$

and E(n) = C − 1 (not an integer). This is called the geometric distribution or the negative binomial distribution.

Multiple Variable Case

For the multivariate case, we can use the appropriate extension of any of these three methods for generating random variables. For example, for acceptance-rejection, if we want to generate x and y according to f(x,y):



needs 3 random numbers, u_1, u_2, u_3 for the two-dimensional case. u_1 and u_2 pick x and y and may be generated uniformly or according to some importance-sampling technique; u_3 is generated uniformly on the appropriate range and is the test variable.

SECTION B. STATISTICAL INFERENCE

Chapter 5. Properties of Estimators

So far, we have considered probability theory. Once we have decided which p.d.f. is appropriate to the problem, we can make direct calculations of the probability of any set of outcomes. Apart from any possible uncertainty about the p.d.f., this is a very straight-forward and mathematically well-defined procedure.

The problem we are now faced with is the inverse of this. We have a certain set of data which have already been sampled from some unknown parent p.d.f. We will try to infer from the observed frequency distribution of the data what this parent p.d.f. looks like. We may have a model restricting the form of this p.d.f. to a certain family with one or more parameters θ_i. In this case, called parametric inference, we want to determine the θ_i. There is another common form of inference, non-parametric inference, in which we admit that we don't know much about the family of the p.d.f. We then concentrate on general statements about the p.d.f., such as its mean, width, or whatever, without assuming anything for its parametrization. In these notes, we will concentrate mainly on the problems of parametric inference, which are certainly formidable.

The study of calculations made using probability is sometimes called direct probability. Statistical inference is sometimes called inverse probability.

In parametric inference, we will usually try to make some sort of statement of the form "$\theta = a \pm b$". The intended interpretation (unless otherwise indicated) is that our *knowledge* of θ is somehow distributed in probability such that our data favor the value "a" for it, but with high probability it could be between a − b and a + b. "b" is often called the "estimated error" and appears on plots of the data as a bar from a − b to a + b. In fact, most people assume that this "knowledge" of θ is a Gaussian of mean a and standard deviation b. The phrase "with high probability" then becomes "with 68.3% probability", and we know precisely how to compute the probability that θ lies in any given interval. Even in more complex cases, for example with unsymmetric error bars, we assume by analogy with the Gaussian that the true value of θ lies within the "one standard deviation" error bars with (roughly) 68.3% probability. This is usually only an approximation; in addition, "two standard deviations" may be considerably different than twice "one standard deviation," if probability levels corresponding to the Gaussian case are intended.

The idea of a probability distribution must here be understood with care. The parameter lies at some exact point; it does not have a p.d.f. except a Dirac δ function (think of saying: "you are 68.3% pregnant"). What is meant is sometimes more properly stated in terms of "confidence". We hope that in a large number of identically-performed measurements, the statistically-determined error intervals (using the same techniques every time) will cover the true value the stated fraction of the time. This probability to which we refer reflects our *confidence* that the true value actually occurs in the given interval.

Finally, we shall see that error bars are difficult to estimate. Just as our best estimate for θ has an error, the error bars have errors. Any confidence intervals which neglect this source of uncertainty are therefore very approximate, and should not be taken too literally. In addition, there are often systematic errors. These are contributions to the measured results which come from sources whose exact effects are unknown. They usually push all the measured data systematically in one direction or another, but they may vary with time or with some of the experimental conditions. After making one's best attempts to estimate and correct for systematic effects there is usually a residual uncertainty which must somehow be incorporated in the confidence intervals. Some experimenters choose to be conservative and quote errors which may be somewhat larger than the true errors. This protects science from drawing overly strong conclusions which might turn out to have been based on an accidental underestimate of the error. However, the value to science of a perfectly valid measurement is reduced when the errors are increased (and the errors may have been accidentally overestimated anyway), so one should exercise discretion about this sort of precautionary increase in the errors.

Suppose we have a set of numbers x_i which are sampled from some population with parameter θ. For example, we measure some quantity n times. θ is the true value of the quantity. The x_i in this example are clustered about θ in some way which depends upon the measuring process. We often assume that the x_i are sampled from a Gaussian of mean θ and width determined by the accuracy of the measurement. We wish to estimate θ. To begin, we construct a function of the x.

Definition: A *statistic* is any function of the observations in a sample which does not depend upon any of the unknown population characteristics.

Example: $\overline{x} = \frac{1}{n}\sum x_i$

If you can calculate it purely from the data sample plus known quantities, it is a statistic.

We want to construct an estimator for θ:

Definition: An *estimator* is any statistic which is going to be announced as our guess at the value of some constant of nature, θ. (θ may be a vector, i.e., represent several unknown quantities.)

We will usually write estimators as: $\hat{\theta}$. Most of the time we will be interested in finding a single point which represents our best estimate of the true value of θ:

Point Estimation

There are certain useful properties which an estimator should have. There are, in general, numerous (for example, ∞) estimators one could construct for any θ. We want to choose the one which is likely to be best. Most of the possible estimators are obviously stupid, but there may be a surprising number which are not. Choosing often involves a trade-off.

Examples of desirable properties (not necessarily in order of importance):

> Unbiased; minimum variance (the variance of an estimator expresses how well it reproduces itself in repreated experiments. If unbiased, it also measures how well it clusters around the true value in repeated measurements); minimum variance (i.e., efficiency); consistent; sufficient; robust (insensitive to errors in our assumptions); minimum loss of information; minimum loss if the estimator is not close to the true value; minimum computer cost, etc.

We will discuss most of these in what follows.

We won't cover information theory here, but there can be cases where information is lost in order to improve the variance; e.g., if $\hat{m} < 0$, where m is known from theory to be positive, we can set $\hat{m} = 0$ and quote an upper limit. The best course is probably to do both — quote $\hat{m}$ as measured and $\hat{m} = 0$. Then people who wish to average experiments, for which one usually needs $\hat{m}$, can do so.

Let us consider estimators of location. If we have n observations of a quantity, e.g.,

mass of a stable particle	— the spread in the data is due to measurement errors alone, e.g., a Gaussian distribution;
lifetime of a particle	— the spread is due to the intrinsic spread in the quantity (i.e., individual decay times), sometimes with additional spread due to measuring error;
mass of a resonance	— the spread is due to the Breit-Wigner (Cauchy) plus a contribution due to the mass resolution (i.e., the error in measuring the mass).

We can suggest several estimators, even for the simple case where we are just interested in location. Each of these might be best under certain circumstances:

For *example*:

$\hat{\mu} = \overline{x} = \frac{1}{n}\sum x_i$ mean — Probably the most heavily used estimator of location. It is also the best under many circumstances, but it can be sensitive to mismeasured data

$\hat{\mu}$ = median — not very sensitive to fluctuations in the tails, or to the presence of heavy-tailed contributions to the distribution; it is therefore more robust than $\overline{x}$ but it has a larger variance than the mean if the data are well-described by a common Gaussian

$\hat{\mu}$ = average of the two extreme values: $[\min(x_i) + \max(x_i)]/2$

— This one is very sensitive to the tails of the distribution, but it may be the best estimator if the underlying distribution is uniform.

$\hat{\mu} = \overline{x}_{\text{trimmed}}$, the trimmed mean: discard the smallest and largest y% (e.g., 10%) and then average.

— This is relatively insensitive to problems in the tails. May not be the "best" estimator if there are no problems in the tails, but it works reasonably well in many real cases.

We will study ways of constructing estimators with reasonable properties. First, let us discuss these properties.

Bias

Since a statistic is a function of random variables, it is itself a random variable. Therefore we can speak of its expectation value. If $E(\hat{\theta}) = \theta$, the true value, the statistic is *unbiased.* To calculate this expectation, we need some assumptions about the underlying distribution.

For example, for the binomial, let us try $\hat{\mu} \equiv \hat{p} = \frac{1}{n}\sum x_i$. This is just k/n, where k = number of successes. Since we assume the binomial, $E(\hat{\mu}) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum p = p$. Therefore $\hat{\mu}$ constructed as above is an unbiased estimator for p.

In general, the sample mean, $\hat{\mu} \equiv \overline{x} = \frac{1}{n}\sum x_i$, is an unbiased estimator of the true mean μ:

$$E(\hat{\mu}) = \frac{1}{n}\sum E(X_i) = \frac{n\mu}{n} = \mu\,,$$

assuming each measurement has the same expectation value μ. That is, over a large number of similar experiments, the average value of $\hat{\mu}$ will be μ.

Now look at the sample mean square.

$S_X^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$. Is this an unbiased estimator for the population variance (i.e., the variance on a single observation, not on the mean)? Assume the population variance is finite.

$$E(S_X^2) = \frac{1}{n}E\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] \qquad \text{(because } \overline{x} \equiv \frac{1}{n}\sum x_i\text{)}$$

$$= \frac{1}{n}\left\{\sum(\sigma^2 + \mu^2) - \frac{1}{n}E[(\sum x_i)^2]\right\} \qquad \text{(see below)}$$

$$= \frac{1}{n}\left\{n\sigma^2 + n\mu^2 - \frac{1}{n}[V(\sum x_i) + [E(\sum x_i)]^2]\right\}.$$

We are assuming each measurement x_i is independent and has the same true mean μ and true variance σ^2, i.e., that it is sampled from the same parent population.

These follow from:

$E(A^2) = V(A) + [E(A)]^2$, (where here $A = \sum x_i$);

$V(\sum x_i) = \sum V(x_i)$ (by independence).

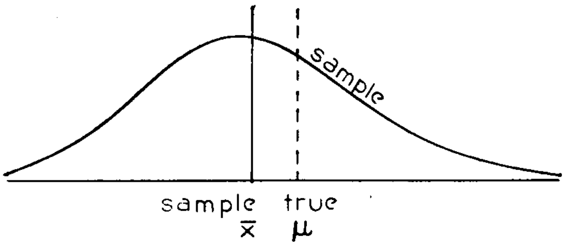
Therefore

$$E(S_X^2) = \frac{1}{n}\left\{n\sigma^2 + n\mu^2 - \frac{1}{n}n\sigma^2 - \frac{1}{n}(n\mu)^2\right\} = \frac{1}{n}(n-1)\sigma^2 \neq \sigma^2\,,$$

and S_X^2 is a biased estimator for σ^2.

The key step here, causing the bias, is that we used $\overline{x}$, the sample mean, in calculating S_X^2. Since this is determined by the sample, it "moves around" a bit to follow the sample.

For example:



The center of mass of the sample distribution is pulled a little to the left by random fluctuations. The average distance (in absolute value) of the x_i from the sample mean is less than from the true mean; the average squared distance is therefore usually also less. Therefore, since the variance is the average squared distance from the true mean, the sample variance is less than the population variance on average (in individual cases it could be larger).

If n is large, this bias is very small, and the sample mean square is often used to estimate the variance because it is convenient. But to get an unbiased estimator, one needs to use

$$S^2 = \left(\frac{n}{n-1}\right) S_X^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2 \quad ,$$

This calculation does *not* depend upon x_i being Gaussian or any other particular distribution. If the x_i's *are* Gaussian, or if n is large enough that the C.L.T. applies, we can see that $\sum z_i^2 = \frac{1}{\sigma^2} \sum (x_i - \overline{x}^2)$ will be a χ^2. We have one relationship among the z_i's:

$$\sum z_i = \frac{1}{\sigma} \sum (x_i - \overline{x}) = 0, \text{ since } \frac{1}{n} \sum x_i = \overline{x} \, ;$$

therefore, $\sum z_i^2$ will be $\chi^2(n - 1)$, since there are only $n - 1$ independent Gaussians in the sum.

$E[\chi^2(n - 1)] = n - 1$, which is another way of seeing that $E(S^2) = E\left(\frac{\sigma^2}{n-1}\chi^2\right) = \sigma^2$. This is of more than passing interest: If we have n measurements x_i of a quantity, with $k \leqslant n$ relationships (*constraints*) among them, then the number of degrees of freedom of the χ^2 constructed from $f[\sum x_i^2]$, where f is the appropriate standardizing function, will be $(n - k)$.

For example: if we constructed our sample mean square as $S_i^2 = \sum (x_i - \mu)^2$, where μ is now the *true* mean rather than something extracted from the data, then S_i^2/n would be an unbiased estimator for σ^2, the population variance.

Variance

Having constructed an estimator, we want to know its variance. Again, we need to make assumptions about the underlying distribution. Having made those assumptions, we can calculate the expected variance of our estimator. If the variance is small, and the estimator unbiased, we can expect the estimator to cluster closely around the true value if the experiment is repeated a large number of times. If the estimator is biased, a small variance means that repeated measurements will yield $\hat{\theta}$'s which cluster tightly around the biased (i.e., wrong) value.

For example, the variance of the sample mean we already know:

$$V\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2}\sum V(x_i) = \frac{\sigma^2}{n}\,.$$

How about the sample variance?

$$V(S^2) = V\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right\}$$

In the case of normal variables or the asymptotic limit where the C.L.T. applies,

$$V(S^2) = V\left[\frac{\sigma^2\chi^2(n-1)}{n-1}\right] = \frac{\sigma^4}{(n-1)^2}V[\chi^2(n-1)]$$

$$= \frac{\sigma^4}{(n-1)^2}\,2(n-1) = \frac{2\sigma^4}{n-1}\,.$$

Thus, S^2, which we can use to estimate the population variance, itself has an error. How can we estimate *this* error if we don't know σ? The usual procedure is to put in our best guess for σ^2:S^2.

$$V(S^2) \cong \frac{2(S^2)^2}{n-1}\,.$$

Caution: Because this is itself an estimate, it is only approximate. If S^2 is too large, then our guess at the error is too large; if too small by the same amount our guess for the error is too *small.* This might (probably would) lead us to deduce that an experiment which measured a low value was more precise than one which measured a high value.

Proposition. The best estimator for a quantity is that unbiased estimator with minimum variance.

This proposition is by no means obvious. In many cases, we might decide to choose an estimator that violates it, because we fear that the assumptions upon which we based our study of bias or variance are questionable, or for some other reason. In many cases, we won't be able to use the best estimator in this sense because the "best" one has some undesirable other properties (see earlier list of desirable properties). We will now discuss some of the implications, and the pros and cons, of this proposition.

There is usually a trade-off of some kind involved. For example, we might be tempted to trade a small bias for a large gain in reducing variance. But this is a risky business. For example, if $\hat{\theta}$ is unbiased, then $\hat{\theta}$f is biased, where f is a number $\neq 1$. $V(f\hat{\theta}) = f^2V(\hat{\theta})$. Therefore, by choosing $f < 1$ we improve the variance by f^2, but introduce a bias. In the limiting case of this, we could simply use as an estimator a fixed number, e.g., 5. If somebody asks our guess as to the value of a parameter, announce "5". This is very clever because the variance = 0 (5 never changes, no matter how many experiments we do). However, except in rare cases, the bias is enormous.

Mean Squared Error

The question of willingness to accept a bias is tricky. Statisticians introduce another figure of merit, the mean squared error, which incorporates any bias into its definition:

Definition. The mean squared error is

$$\text{MSE} = E[(\hat{\theta} - \theta)^2] \quad \text{(calculating this requires knowing } \theta \text{, for the moment)}$$

$$= V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \ .$$

Here, $E(\hat{\theta}) - \theta$ is the *bias* $b_\theta(\hat{\theta})$; bias is always the *expected* bias, which may or many not be a function of θ. This is just an expression of what we said before. If the bias = 0, estimate the error in our estimator just by looking at its variance.

We could use the MSE as a criterion for selecting estimators. In some cases, we might choose a biased $\hat{\theta}$ with smaller MSE than a certain unbiased θ.

For example, look at estimators for the variance σ^2 of a normal population. Try the form $Z^2 = A \sum (x_i - \overline{x})^2$ and find the constant A such that the MSE is minimized.

$$\text{MSE} = V(Z^2) + b^2 \ ; \quad b = E(Z^2) - \sigma^2 \ .$$

$$V(Z^2) = V[A(n-1)S^2] = 2A^2(n-1)\sigma^4 \ , \quad \begin{array}{l}\text{from the analysis of } S^2 \text{ above} \\ \text{(for normal } x_i) \ ,\end{array}$$

$$E(Z^2) = E[A(n-1)S^2] = A(n-1)\sigma^2 \ .$$

Therefore,

$$\text{MSE} = \sigma^4[2A^2(n-1) + A^2(n-1)^2 - 2A(n-1) + 1] \ .$$

Now we can minimize:

$$\frac{\partial \text{MSE}}{\partial A} = 0 \qquad \text{requires } A = \frac{1}{n+1} \ .$$

Therefore, even with the bias, we can improve our mean squared error if we estimate σ^2 by $\hat{\sigma}^2 = \frac{1}{n+1}\sum(x_i - \overline{x})^2$. This comes about because $\sum(x_i - \overline{x})^2$ is proportional to a χ^2, which is asymmetrical. Although any particular measurement is more likely to be below the mean than above (i.e., the mean lies above the median), if it is above, it may be far above (the χ^2 has fairly long tails toward high values). Since large values weigh heavily in averages of squared errors, we can reduce the MSE somewhat by deliberately reducing our estimate slightly in all cases. As n increases, the χ^2 becomes more nearly symmetrical, and also $\frac{1}{n+1}$ approaches $\frac{1}{n-1}$ and the bias disappears.

Should we do this? People are sometimes willing to accept a biased estimator for a number of reasons. One of the most frequent is convenience. But we always insist that any bias be small compared with the other errors in the problem. It is not worth going to a lot of trouble refining an estimate if the refinement results in a correction which is much less, for example, than one's uncertainties in the systematic errors. If we accept the principle that the bias must be small, then the improvement in the MSE by taking a biased estimator is probably also small, and it may not be worth *introducing* a small bias to get such an improvement in the MSE. If the bias is difficult to get rid of, as in some maximum likelihood problems, that is a different matter. In any case, one must exercise judgment based upon the specific problem. Often, a Monte Carlo investigation is the only way to study the situation.

One important reason for avoiding bias where possible is that it must be possible to combine the results of different experiments in some sort of average. This can only be done with some reliability if the experiments quote the same sort of result, which usually means *unbiased.* Averaging experiments is somewhat like going to higher n (collecting more data), except that some biases disappear at large n, as we have seen and will discuss further shortly. However, if we average, say, five experiments of similar size and bias, the bias remains in the average, which otherwise appears to be much more precise than any of the individual experiments. One could, of course, remove the bias at the averaging stage if the bias is known.

Another problem with using MSE as sole figure of merit in choosing among estimators is that if the distributions of two estimators are of different shapes, the MSE may not be a fair basis for comparison. For example, it may be that an estimator with infinite MSE may be acceptable if the divergence is produced by an effect of very small probability.

Still another problem with using the MSE as the sole judge of estimators is that there may be no estimator with *uniformly* better MSE. For example, let Y be the number of successes in n Bernoulli trials, with probability p of a success on each trial. We wish to estimate p. Consider the two estimators:

$$\hat{p}_1 = Y/n$$

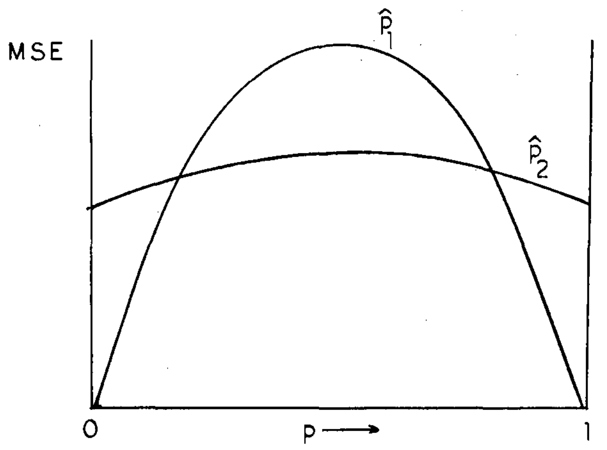
$$\hat{p}_2 = (Y + 1)/(n + 2).$$

We can compare the MSE's of these two estimators without actually knowing their p.d.f.'s, since we know that V(Y) = npq (q = 1 − p) and E(Y) = np. Then

$$\mathrm{MSE}(\hat{p}_1) = pq/n\ ;$$

$$\mathrm{MSE}(\hat{p}_2) = \left(\frac{1}{n + 2}\right)^2 [npq + 1].$$

For n = 4, the MSE's look as follows:



To choose the one with smallest MSE, we would need to have some pre-conceived notion as to the likely range of p. For example, if we expected $p \cong 1/2$ from theory or earlier experiments, we might be led to choose $\hat{p}_2$. This is a Bayesian decision. However, we should consider other facts, as well. Most importantly, $\hat{p}_2$ is biased, $\hat{p}_1$ is not. Most physicists would normally prefer $\hat{p}_1$ for that reason, unless some compelling reason otherwise should present itself.

Some problems with unbiased estimators, other than the fact that they may not be convenient, are:

a) Unbiased estimators may not exist. If so, that's a real problem for someone who always insists on unbiasedness in estimators.

b) They may not have the smallest MSE.

c) The property of unbiasedness is not necessarily invariant under change of variables. For example, S^2 is unbiased for σ^2, but S not unbiased as an estimator for σ:

An unbiased estimate for σ, looking like $\sqrt{S^2}$, is

$$\hat{\sigma}_1 = \frac{\sqrt{n}\,\Gamma[\frac{1}{2}(n-1)]}{\sqrt{2}\,\Gamma(\frac{1}{2}n)} \sqrt{\frac{1}{n}\Sigma(x_i - \hat{x})^2}\,;$$

recall that, for m an integer, $\Gamma(m+1) = m!$ and $\Gamma(m+1/2) = \frac{(2m-1)!!}{2^m}\sqrt{\pi}$..

Then it may be shown that $E(\hat{\sigma}_1) = \sigma$ if the distribution of x_i is normal. For $n = 1$, no distribution whose width we could estimate exists. For $n = 2$, $\hat{\sigma}_1 = 1.77 \times$ RMS, where RMS is the root-mean square width of the sample; the RMS is given by the radical. As n increases, $\hat{\sigma}_1$ converges from above to the RMS. For $n = 10$, $\hat{\sigma}_1$ is less than 10% more than the RMS.

An old dispute (1920's) concerned whether or not $\hat{\sigma}_1$ above was better, or

$$\hat{\sigma}_2 = \sqrt{\frac{\pi n}{2(n-1)}} \cdot \frac{1}{n} \Sigma|x_i - \hat{x}|\,.$$

Also, $E(\hat{\sigma}_2) = \sigma$ in the normal case.

It turns out that $\hat{\sigma}_1$ has uniformly better MSE than $\hat{\sigma}_2$, if normality is satisfied. However, if there are departures from normality, $\hat{\sigma}_2$ can be much better, because it doesn't weight the tails as heavily (it is therefore more *robust*).

Consistency

When we take more data, we expect our ability to estimate parameters θ to improve, provided the data is relevant. If the estimator is biased, it will converge to the biased value, unless the bias disappears as the data sample increases. If the bias does not vanish in the asymptotic limit, either we agree to change to a more sophisticated estimator if we get more data, or we should use an estimator from the outset which is *consistent*.

Definition: An estimator $\hat{\theta}$ for θ is said to be *consistent* if

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| \geqslant \epsilon) = 0 \text{ for any } \epsilon > 0 .$$

Clearly, if $\hat{\theta}$ is an average of the data, then, by the law of large numbers (when it applies), $\hat{\theta}$ is a consistent estimator for the population mean:

$$P\left(\left|\frac{1}{n} \sum x_i - \mu\right| \geqslant \epsilon \right) \leqslant \frac{1}{\epsilon^2} \frac{\sigma^2}{n} .$$

Now we can drag out our favorite counter-example, the Cauchy: $z_n = \frac{1}{n} \sum x_i$ is *not* a consistent estimator for the population parameter μ.

$$C(x;\mu,\alpha) = \frac{1}{\pi\alpha} \frac{1}{1 + (x - \mu)^2/\alpha^2}$$

We have earlier shown that the distribution of z_n is the same as that of x: $C(z_n ; \mu,\alpha)$, i.e., there is no convergence to anything at all, much less to μ.

The biased estimators we have looked at so far are consistent, e.g.,

$$\frac{1}{n+1} \sum(x_i - \hat{x})^2 \cong \frac{1}{n-1} \sum(x_i - \hat{x})^2 \quad \text{as } n \to \infty .$$

Next, we want to investigate the conditions under which we might find a uniformly minimum variance estimator, and how we might recognize one if we have one.

Chapter 6. The Rao-Cramér-Frechet Bound

It turns out that there is a lower limit to the variance of an estimator under certain general conditions. Therefore we can recognize when we have found an estimator of the smallest possible variance.

Theorem. Rao-Cramér-Frechet Inequality

Let $X_1, \dots, X_n$ be sampled from some population $f(x,\theta)$, θ a parameter. θ is whatever is being estimated, which might be a function of some fundamental parameter of the p.d.f. Note that the X_i are independent and identically distributed. Assume the range of x, $\Omega = \{x : f(x,\theta) > 0\}$, does not depend upon θ. Choose an estimator $\hat{\theta}$ for θ, with bias $b_{\hat{\theta}}(\theta) = E(\hat{\theta}) - \theta$, such that

$V(\hat{\theta}) < \infty$ For all θ. Then

$$V(\hat{\theta}) \geqslant \frac{\left[1 + \frac{\partial b}{\partial \theta}\right]^2}{I(\theta)} \quad ; \quad \hat{\theta} = \hat{\theta}(x_1, \dots, x_n) ,$$

where $I(\theta)$ is called the Fisher information number.

The Fisher information number is given by

$$I(\theta) = E\{[\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ell n\ f(x_i,\theta)]^2\}$$

Note: $\frac{\partial}{\partial \theta} \ell n\ f(x,\theta)$ is defined as the *score* function, written $S(x,\theta)$.

Since the score is a function of a r.v., it is a r.v. The score function is very important for maximum likelihood analysis (Chapter 10).

Proof

We will prove this theorem for continuous variables. First, we assume that we can interchange the order of integration and differentiation for the expectation of any statistic T(x) of finite variance:

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} T(x) f(x,\theta) dx = \int_{-\infty}^{\infty} T(x) \frac{\partial}{\partial \theta} f(x,\theta) dx .$$

This is called the regularity condition; it is quite generally true for statistics of physical interest. Then, if we set $T(x) = 1$,

$$\int 1 \cdot \frac{\partial}{\partial \theta} f(x,\theta) dx = \frac{\partial}{\partial \theta} \int f(x,\theta) dx = 0 ,$$

since $f(x,\theta)$ is normalized for all θ, and therefore

$$E\left[\frac{\partial}{\partial \theta} \ell n\ f(x,\theta)\right] = E[S(x,\theta)] = \int \left[\frac{\partial}{\partial \theta} \ell n\ f(x,\theta)\right] f(x,\theta) dx$$

$$= \int \frac{1}{f(x,\theta)} \left[\frac{\partial}{\partial \theta} f(x,\theta)\right] f(x,\theta) dx$$

$$= \int \frac{\partial}{\partial\theta} f(x,\theta)dx = \frac{\partial}{\partial\theta}\int f(x,\theta)dx = 0 .$$

Therefore

$$E\left\{\left[\frac{\partial}{\partial\theta}\sum_i \ell n\, f(x_i,\theta)\right]^2\right\} = E\left\{\left[\sum_i \frac{\partial}{\partial\theta}\ell n\, f(x_i,\theta)\right]^2\right\}$$

$$= V\left[\sum_i \frac{\partial}{\partial\theta}\ell n\, f(x_i,\theta)\right] + \left\{E\left[\sum_i \frac{\partial}{\partial\theta}\ell n\, f(x_i,\theta)\right]\right\}^2 , \text{ since E is a linear operator}$$

$$= V\left[\sum_i \frac{\partial}{\partial\theta}\ell n\, f(x_i,\theta)\right]$$

$$= \sum_i V\left[\frac{\partial}{\partial\theta}\ell n\, f(x_i,\theta)\right], \text{ since the } x_i \text{ are all independent}$$

$$= nV\left[\frac{\partial}{\partial\theta}\ell n\, f(x,\theta)\right], \text{ since the } x_i \text{ are identically distributed.}$$

$$= nI_1(\theta)\,; \qquad I_1 = \text{information from one observation.}$$

The last equality follows from the same sequence of steps being applied to $I_1(\theta)$. The Fisher information contained in n independent, identically distributed measurements is n times the information contained in any one. Similarly,

$$E\left[\underbrace{\hat\theta\frac{\partial}{\partial\theta}\sum_i \ell n\, f(x_i,\theta)}_{S}\right] = \int\cdots\int\hat\theta\left[\frac{\partial}{\partial\theta}\ell n\,\prod_i f(x_i,\theta)\right]\prod_i[f(x_i,\theta)dx_i]$$

$\hat\theta$ is a function of all the x_i's; hence we have to look at a multiple integral here.

$$= \int\cdots\int\hat\theta\left[\frac{1}{\Pi f(x_i,\theta)}\,\frac{\partial}{\partial\theta}\Pi f(x_i,\theta)\right]\Pi[f(x_i,\theta)dx_i]$$

$$= \int\cdots\int\hat\theta\frac{\partial}{\partial\theta}[\Pi f(x_i,\theta)]\,dx_i\cdots dx_n$$

$$= \frac{\partial}{\partial\theta}\int\cdots\int\hat\theta f(x_i,\theta)\cdots f(x_n,\theta)dx_1\cdots dx_n$$

$\hat\theta$ is a statistic and therefore does not depend upon θ

$$= \frac{\partial}{\partial\theta}[\theta + b_{\hat\theta}(\theta)] = 1 + \frac{\partial}{\partial\theta}b(\theta) \ .$$

Now, we know that both $\hat\theta$ and the quantity $\sum\limits_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)$ are r.v.'s, and we can calculate their correlation:

$$\mathrm{COV}\left[\sum_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta),\hat\theta(x_i)\right] \qquad \text{Covariance between the score and the estimator.}$$

$$= E\left[\sum_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)\cdot\hat\theta(x_i)\right] - E\left[\sum_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)\right]E[\hat\theta(x_i)]$$

$$= E\left[\hat\theta\sum_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)\right] = \left[1 + \frac{\partial}{\partial\theta}b\right] .$$

Therefore the correlation coefficient is given by

$$\rho^2 = \frac{\left\{\mathrm{COV}\left[\sum\limits_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta),\hat\theta(x_i)\right]\right\}^2}{V\left[\sum\limits_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)\right]V[\hat\theta(x_i)]} \text{ which must be} \leqslant 1\ ,$$

and therefore

$$\frac{\left[1 + \frac{\partial}{\partial\theta}b\right]^2}{V\left[\sum\limits_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)\right]} = \frac{\left[1 + \frac{\partial}{\partial\theta}b\right]^2}{E\left\{\left[\sum\limits_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)\right]^2\right\}} \leqslant V(\hat\theta)\ . \quad \text{Q.E.D.}$$

This is called the Rao-Cramér-Frechet Inequality (perhaps the more usual name is the Rao-Cramér Inequality, but Frechet actually had precedence).

Since we don't know the bias *a priori,* we would have to search all possible biased estimators to find the best one, and the inequality is not as useful as it might be: we don't know when to stop looking. In particular, note that the bias of a constant C is C − θ, and therefore the RCF bound = 0. That's not a useful lower bound for a variance, a positive quantity. But if we restrict ourselves to the class of unbiased estimators, we have a very useful bound:

$$\sigma^2(\hat\theta) \geqslant \frac{1}{E\left\{\left[\sum\limits_i \frac{\partial}{\partial\theta}\ell n\ f(x_i,\theta)\right]^2\right\}} = \sigma^2_{\min}\ .$$

For the multi-variate case (i.e., $\hat\theta$ is a vector of more than one component), we write an information matrix:

$$\text{e.g., } I_{11} = E\left\{\left[\sum_i \frac{\partial}{\partial\theta_1}\ell n\ f(x_i,\theta)\right]^2\right\}.$$

In general,

$$I_{\ell m} = E\left\{\left[\sum_i \frac{\partial}{\partial\theta_\ell}\ell n\ f(x_i,\theta)\right]\left[\sum_i \frac{\partial}{\partial\theta_m}\ell n\ f(x_i,\theta)\right]\right\}.$$

Then we could use

$$\sigma^2(\hat\theta_1) \geqslant \frac{1}{I_{11}(\theta)} \ ,$$

but a stronger (higher) limit is

$$\sigma^2(\hat\theta_1) \geqslant [I(\theta)]_{11}^{-1}\ , \text{ the } (1,1)^{\text{th}} \text{ element of } I^{-1}\ ,$$

which takes account of the correlations between the variables.

For unbiased estimators, we define the *efficiency* of an estimator $\hat\theta$:

$$\epsilon(\hat\theta) = \frac{\sigma^2_{\min}}{\sigma^2(\hat\theta)} \leqslant 1\ .$$

This is a generally useful quantity, although it may well be that an unbiased estimator of efficiency 1 may not exist for a particular problem. Or it may be efficient only if the true value of θ lies in some limited range, not for others—i.e., not uniformly efficient.

In terms of the MSE,

$$\text{MSE}(\hat\theta) = V(\hat\theta) + b_{\hat\theta}^2(\theta) \geqslant \frac{[1 + \frac{\partial}{\partial\theta}\, b_{\hat\theta}(\theta)]^2}{I(\theta)} + b_{\hat\theta}^2(\theta) \quad \text{(biased case)}$$

$$\text{and MSE } (\hat\theta) = V(\hat\theta) \geqslant \frac{1}{I(\theta)} \quad \text{(unbiased case)}\ .$$

Example

$$\text{Let } f(x, \theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-\frac{x^2}{2\theta}} \ ; \ \theta \equiv \sigma^2$$

i.e., $N(x; 0, \sqrt{\theta})$, the mean is known. (We can also do this with vector θ, e.g., $\vec\theta = (\mu,\sigma^2)$, similarly. This will be touched upon later.) Now,

$$\ell n\ f = -\frac{1}{2}\ell n(2\pi) - \frac{1}{2}\ell n\ \theta - \frac{x^2}{2\theta^2}\ .$$

The score $= \frac{\partial}{\partial\theta}\ell n\ f = -\frac{1}{2\theta} + \frac{x^2}{2\theta^2}$. Therefore, $I_1(\theta) \equiv$ information in a single observation

$$= E\left\{\left[\frac{\partial}{\partial\theta}\ell n\ f\right]^2\right\} = E\left\{\left[-\frac{1}{2\theta} + \frac{x^2}{2\theta^2}\right]^2\right\}$$

$$= \frac{E[(x^2 - \theta)^2]}{4\theta^4} = \frac{V(X^2)}{4\theta^4} = \frac{1}{2\theta^2}. \qquad \text{Exercise for the reader}$$

Now note that with μ *known* (even if not 0), the estimator:

$\hat{\theta} = \frac{1}{n}\sum(x_i - \mu)^2$ has $E(\hat{\theta}) = \sigma^2$, and therefore

$$V(\hat{\theta}) = \frac{1}{n}V[(X - \mu)^2] = \frac{1}{n}\{E[(X - \mu)^4] - (E[(X - \mu)^2])^2\}$$

$$= \frac{2\theta^2}{n} \geqslant \frac{1}{nI_1(\theta)} = \frac{2\theta^2}{n}.$$

Therefore $\hat{\theta} = \frac{1}{n}\sum(x_i - \mu)^2$ is efficient. Note that, in these calculations, the true θ appears. We make these calculations as if we *knew* θ; then we know the conclusions hold even when we don't.

An alternative formulation of the RCF bound:

Let $S = \text{score} = \frac{\partial}{\partial\theta}\ell n\, f(x,\theta)$.

Then $I(\theta) = E\left[(\frac{\partial}{\partial\theta}\ell n\, f)^2\right] = E[S^2]$.

for one observation

$$= - E\left[\frac{\partial^2}{\partial\theta^2}\ell n\, f(x,\theta)\right] = - E\left[\frac{\partial S}{\partial\theta}\right].$$

The proof is left as an exercise. For more than one variable, the (ℓ, m)th element of the information matrix becomes

$$I_{\ell m}(\theta) = - E\left[\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\ell n\, f(x,\theta)\right].$$

Let us look at a distribution other than the normal. The normal often hides important points because of its symmetry.

Example. The exponential.

Let $X_1,..., X_n$ be independent, identically distributed from the exponential, $f(x, \mu) = \frac{1}{\mu}e^{-x/\mu}$;$\mu \geqslant 0$. We must estimate μ. Then

S_j = score of jth value

$$= \frac{\partial}{\partial\mu}[-\log\mu - x_j/\mu] = - \frac{1}{\mu} + \frac{x_j}{\mu^2}.$$

Therefore

$$I_j(\mu) = -E\left[\frac{\partial S_j}{\partial \mu}\right] = -E\left[\frac{1}{\mu^2}\left(1 - \frac{2x_j}{\mu}\right)\right]$$

$$= -\frac{1}{\mu^2}\left(1 - 2\,\frac{\mu}{\mu}\right) = \frac{1}{\mu^2}\,.$$

Therefore $I(\mu) = nI_1(\mu) = \frac{n}{\mu^2}$, requiring $\sigma^2(\hat{\mu}) \geqslant \frac{\mu^2}{n}$ (if unbiased).

Notice that, so far, we haven't talked about any specific estimator. Try $\hat{\mu} = \frac{1}{n}\sum x_i = \overline{x}$.

$$E(\overline{x}) = \mu \text{ ; unbiased.}$$

$$V(\overline{x}) = \frac{n}{n^2}V(x) = \frac{1}{n}\left[\int_0^\infty x^2\,\frac{1}{\mu}\,e^{-x/\mu}dx - \mu^2\right]$$

$$= \mu^2/n\,.$$

Therefore $\hat{\mu} = \overline{x}$ is an efficient estimator for μ.

The full score function is

$$S = \sum_i S_i = -\frac{n}{\mu} + \frac{\sum x_i}{\mu^2}\,.$$

Note that our estimator is a linear function of the score:

$$\hat{\mu} = \frac{1}{n}\sum x_i = \frac{\mu^2}{n}S + \mu\,.$$

This is not a coincidence, as we show now.

Efficient Estimators

The R.C.F. bound was obtained from

$$\rho^2 = \frac{[COV(S,\hat{\theta})]^2}{V(S)\,V(\hat{\theta})} \leqslant 1$$

We have seen that $\rho = \pm 1$ corresponds to a straight line. Therefore we can construct an efficient estimator, under the conditions of the theorem, if and only if $\hat{\theta}$ is a linear function of S, with probability 1. This can be violated at isolated points whose probability is zero.

Now write $A'(\theta)$, $B'(\theta)$ as the derivatives of some functions of θ, but not functions of x (Note: the covariance in ρ above is with respect to x).

Then, for $\hat{\theta}$ an unbiased efficient estimator,

$$S = A'(\theta)\hat{\theta} + B'(\theta) = \frac{\partial}{\partial\theta}\,\ell n\, f(x,\theta).$$

Then $\ell n\ f(x,\theta) = A(\theta)\hat{\theta} + B(\theta) + K(x)$,

where the arbitrary constant K may depend upon x, but not θ. Remember that $\hat{\theta}$ is a statistic and therefore not a function of θ; $\hat{\theta} = \hat{\theta}(x)$.

Then
$f(x,\theta) = \exp\{A(\theta)\hat{\theta} + B(\theta) + K(x)\}$ over some interval which is independent of θ.

The normalization is included inside the exponential, usually in $B(\theta)$ and/or $K(x)$.

The Exponential Family

Any p.d.f. of this form is said to belong to the *exponential family.* Thus, in order to find an efficient estimator, the p.d.f. for x must belong to the exponential family, and the estimator must enter it in a certain way.

In general the exponential family is of the form

$f(x,\theta) = \exp\{A(\theta)\ C(x) + B(\theta) + K(x)\}$

Thus only if $\hat{\theta} = C(x)$ is $\hat{\theta}$ efficient.

The converse is also true. If $f(x,\theta)$ is of the above form, with $\hat{\theta} = C(x)$, then $\hat{\theta}$ will be efficient. The product $A(\theta)\ C(x)$ can often be factored into $A(\theta)$ and $C(x)$ in more than one way. For example, let

$$A(\theta)C(x) = Q(\theta)\sum_{i=1}^{n} x_i\ .$$

Then one could take $A(\theta) = Q(\theta)$, $C(x) = \sum x_i$, or, alternatively, $A(\theta) = nQ(\theta)$ and $C(x) = \frac{1}{n}\sum x_i$. Thus, in general, A,B,C, and K are not uniquely specified for a given p.d.f., and o ne may be able to find efficient estimators for several different quantities.

We may have x be a vector; e.g., a number of data sampled from a certain exponential family. Thus, if

$\vec{x} = (x_1,\ldots, x_n)$ are independent, identically distributed from an exponential family, then

$p(\vec{x}, \theta) = \Pi f(x_i,\theta)$ is an exponential p.d.f. in n dimensions.

We may also have θ multidimensional.

For example,

$$N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

$$= \exp\left\{\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left[\frac{\mu^2}{\sigma^2} + \ell n(2\pi\sigma^2)\right]\right\}\ .$$

Then $\vec{\theta} = (\mu,\sigma)$, and A and C are vectors, and $A(\vec{\theta})C(x)$ is a vector dot product:

$A_1(\vec{\theta}) = \mu/\sigma^2$ $C_1(x) = x$

$$A_2(\vec{\theta}) = -\frac{1}{2\sigma^2} \qquad C_2(x) = x^2$$

$$B(\theta) = -\frac{1}{2}\left[\frac{\mu^2}{\sigma^2} + \ell n(2\pi\sigma^2)\right]$$

$$K(x) = 0$$

If either μ or σ is fixed, i.e., known at the outset of the problem, then the normal can be regarded as a 1-parameter distribution belonging to the exponential family.

Discrete distributions can belong to exponential families, also. For example, the Binomial:

$$f(x,\theta) = \binom{n}{x} \theta^x(1 - \theta)^{n-x}; \text{ n fixed}$$

$$= \exp\left\{x\, \ell n\left(\frac{\theta}{1-\theta}\right) + n\, \ell n(1 - \theta) + \ell n\binom{n}{x}\right\}$$

$$A(\theta) = \ell n\left(\frac{\theta}{1-\theta}\right) \qquad C(x) = x$$

$$B(\theta) = n\, \ell n(1 - \theta) \qquad K(x) = \ell n\binom{n}{x}$$

Many others also belong to this family:

Bernoulli: $p^x(1 - p)^{1-x}$

Geometric: $p(1 - p)^x$

The geometric distribution describes the number of failures x before the first success in a Bernoulli process.

Negative Binomial: $\binom{r + x - 1}{x} p^r(1 - p)^x$;

The negative binomial distribution describes the total number of trials x necessary to reach exactly r successes in a Bernoulli process.

Poisson: $e^{-\mu}\dfrac{\mu^x}{x!}$

Exponential: $\dfrac{1}{\mu}e^{-x/\mu}$

Gamma: $\lambda^n x^{n-1}\dfrac{e^{-\lambda x}}{(n-1)!}$

Rayleigh: $\dfrac{x}{\theta^2}\exp\left[-\dfrac{x^2}{2\theta^2}\right]$

All of these examples are one-parameter distributions.

Chapter 7. Sufficient Statistics

Under certain circumstances, it is possible to start with any old unbiased estimator and improve it with one step to the best we can do (this means to the estimator with least possible variance, but not necessarily one that achieves the R.C.F. lower bound). This works at least for the exponential family, where the estimator is going to be some function, not necessarily linear, of C(x) (if it is not linear, it can't be efficient, but that may be the best we can do).

Example

The Poisson. The No-Count Probability. The probability of zero counts in a certain sample or sub-sample is given by

$$\theta = e^{-\mu}, \text{ since}$$

$$\frac{\mu^x}{x!} = 1 \quad \text{if } x = 0\,.$$

We wish to find the best possible estimator for θ, this probability. Assume that we have n categories, which may be bins in a histogram, and further that each bin is filled by a Poisson process with parameter μ, common to each bin. We now observe x_i events in the i^th bin. It can be shown that $\overline{x} = \frac{1}{n}\sum x_i$ is the best estimator for μ, given a random sample of x_i's. We are tempted, therefore, to use $\hat{\theta}_1 = e^{-\overline{x}}$. We have no guarantee that $\hat{\theta}_1$ is unbiased, or minimum variance. In fact, it is neither.

In the discussion about Monte Carlo techniques, we saw that it is sometimes possible to pick the particular member of the *family* of distributions first, with one random number, then to pick *from* that family member second, with a second random number. Here, we just want to know which member of the family best describes the data, essentially the reverse of the first step. Nature has done the second step for us. So we will pick a statistic which is sufficient to describe the family. Then we will use it to construct a best estimator for $e^{-\mu}$.

Before we do that, let us construct some unbiased estimator for θ. Look at our sample x_i, i = 1, ..., n. X_i is the number of counts in the i^th bin. We can make an unbiased estimator for θ if we look just at the first of the x_i, namely x_1. If $x_1 = 0$, set our estimator = 1. If $x_1 \neq 0$, set our estimator = 0. The quantity we are estimating, θ, is, after all, a probability. That is , let T be an estimator for θ, where

$$T = \begin{cases} 1 \text{ if } X_1 = 0 \\ 0 \text{ if } X_1 \neq 0 \end{cases}$$

You don't always need an algebraic function to specify an estimator; sometimes just a table of values.

We estimate the probability of zero events in every bin to be certainty if just $x_1 = 0$; otherwise, we throw up our hands and estimate that there is zero probability that any bin will be empty. This is not very sophisticated, since we have *some* events, and they have to go somewhere, even if the first bin happens to be empty. Is T biased?

$$E(T) = \sum_{x=0}^{\infty} Te^{-\mu}\frac{\mu^x}{x!} = e^{-\mu} + \sum_{x=1}^{\infty} 0 \cdot e^{-\mu}\frac{\mu^x}{x!} = e^{-\mu}\text{, inserting for T.}$$

Therefore T is unbiased, and will give the correct value on the average over a large number of experiments.

$$V(T) = \sum_{x=0}^{\infty} T^2 e^{-\mu} \frac{\mu^x}{x!} - e^{-2\mu} = e^{-\mu} - e^{-2\mu} = e^{-\mu}(1 - e^{-\mu}) \, .$$

We are surely do better than this, since we have neglected all of the information contained in $x_2,...,x_n$, whose values are known. Looking at n bins tells us as much as n whole experiments in which we look only at the first bin. The next step is to find a reduction of the data to as simple as possible a statistic V which contains *all* the information about θ which can be obtained from the sample. That is, given a statistic V (which may be multi-dimensional) fixed at a certain value, the distribution of any other function of the random variables does not then depend upon θ. In particular, $f(\vec{x}|V)$ no longer depends upon θ. Since $f(\vec{x}|V)$ doesn't depend upon θ, we can't extract anything about θ from it. Therefore all the information that is to be had about θ is in V. There are lots of possible choices of V, because any $1-1$ function of V, such as V^3, etc., satisfies the requirement: fix it, and we remove all θ dependence from the distribution. The same is true for any multi-dimensional extension of V, U = (V,Q,R,S,...), where Q,R,S,... denote other statistics. Obviously, one prefers to work with the greatest possible reduction of the data which will satisfy the conditions, and therefore dispense with the superfluous Q,R,S,....

V itself may not be an estimator for the unknown parameters, but it seems clear that we want some function of V for our estimator.

If V satisfies the above requirements, then V is *sufficient.* It is also a statistic, by construction. If V is the greatest possible reduction of the data, it is *minimal sufficient.* Note: In the Poisson example, T is not *sufficient,* but we were able to construct an unbiased estimator for θ using it, anyway.

If V is sufficient, then $f(\vec{x};\theta)$ can be factored: $f(\vec{x};\theta)d\vec{x} = f(\vec{x}|v)f_1(v)|J|d\vec{y}dv$ where $f(\vec{x}|v)|J|$ is a function which doesn't depend upon θ; $f_1(v)$ depends upon θ, but its $\vec{x}$-dependence occurs only through v:

$$f(\vec{x};\theta)d\vec{x} = \underbrace{h(\vec{x})}_{f(\vec{x}|v)|J|} \quad \underbrace{g(v,\theta)}_{f_1(v)}d\vec{y}dv \ .$$

We represent by $d\vec{y}$ the (n − 1)-dimensional differential left after v (here, 1-dimensional) is extracted. This is a change of variables, e.g., from x_1, ..., x_n to $x_1,...,x_{n-1}$, v, or some other change of variables from $(\vec{x})$ to $(\vec{y},v)$, where v is one of the new ones. $|J|$ is the Jacobean for this change of variables. Remember $|J|$ is not needed for discrete distributions.

The choice of V will depend upon the distribution f which we assume describes the data. If this model is accepted, then two sets of data which result in the same value for V will lead to identical conclusions about θ.

Selecting a sufficient statistic V is an example of *data reduction.* There may be a vast amount of information in the data; we may be able to extract all the information about θ that the sample contains by calculating one or a very few numbers, e.g., a mean. Just as in the Monte Carlo case, once the value of V is known, the family member is specified and the rest of the information in the sample may be regarded as having been generated by some random mechanism which has no dependence upon θ. This (sometimes very large) amount of information which is not needed to determine θ has been rendered superfluous by the power of our assumption, that f is the proper p.d.f. After θ has been determined, the actual data points could be used if desired to perform a test of this assumption.

The above factorization of the joint p.d.f. for all the x's is a necessary and sufficient condition for V to be a sufficient statistic. Therefore it may be used to select and/or verify a sufficient statistic.

In most cases, one is primarily interested in minimal sufficient statistics. In our no-count example, we can use the number of events $v = \sum x_i$ (or any $1 - 1$ function of that). Let us show that this v is sufficient:

We know that the Poisson is *reproductive.*

Therefore,

$$V = \sum x_i \text{ is } P_{n\mu}(v) = \frac{e^{-n\mu}(n\mu)^v}{v!} = f_1(v)\,,$$

since each x_i has the same Poisson parameter, μ. Now,

$$f(\vec{x};\mu) = \prod_i e^{-\mu}\,\frac{\mu^{x_i}}{x_i!} = e^{-n\mu}\,\frac{\mu^{(\sum x_i)}}{\Pi x_i!} = \frac{e^{-n\mu}\,\mu^v}{\Pi x_i!}\,.$$

Therefore,

$$\underset{\Big\uparrow}{f(\vec{x}|v)} = \frac{f(\vec{x};\mu)}{f_1(v)} = \frac{e^{-n\mu}\,\mu^v}{\Pi x_i!} \times \frac{v!}{e^{-n\mu}(n\mu)^v}$$

function of n−1
variables, since
V is fixed at v

$$= \frac{v!}{(\Pi x_i!)n^v}\,, \quad \text{(no } |J| \text{ in the discrete case)}$$

which does not depend upon μ, with V fixed at v.

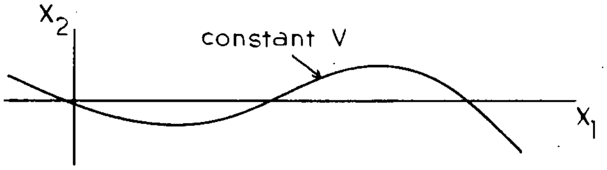
Therefore, V is sufficient for μ. The total number of events is not an estimator for μ but it is all we need to know to construct an estimator for μ, or any function of μ, such as $e^{-\mu}$. Now what?

The trick is to take the expectation of the unbiased estimator T at a fixed value of the sufficient statistic V. This gives a new r.v. which we will show combines the unbiasedness of T with the information content of V:

$$U = E(T|V) = \int_V T f(\vec{x}|V)|J|d\vec{x} \quad \text{(continuous case).}$$

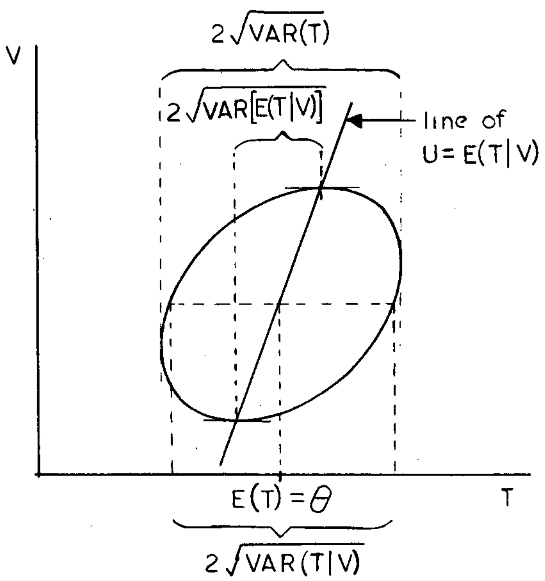
The integral over all $\vec{x}$ means that all $\vec{x}$-dependence is removed from U except that contained in V: if we have a different V, we will get a different U. But U is a function only of that one variable V, and therefore represents a considerable reduction in the complexity of the problem. It is the fact that V is *sufficient* that assures us that U is a statistic, i.e., does not depend upon θ. This is because $f(\vec{x}|V)$ does not depend upon θ; none of the other terms in the integral do, either.

In two dimensions, the integral over all space reduces to the integral over the line of constant V:



In general, this is an (n−1) dimensional integral, for one-dimensional V.

If V is a sufficient statistic, it contains all of the information about θ that is available from the sample. If T is an unbiased estimator, it contains at least some of that same information. Therefore the two statistics T and V must be correlated. If, for the purposes of illustration, T and V are sampled from a bivariate normal, then:



If T is a function of V only, T = T(V), then the correlation is 100%, T is itself sufficient, and U = T, since to fix V at its measured value is to fix T at its measured value.

U is an unbiased estimator, since calculating its expectation value requires integration over all V; for continuous variables we have:

$$E(U) = E[E(T|V)] = \int\left[\int_V Tf(\vec{x}|v)|J|d\vec{x}\right]f_1(v)dv;$$

but $f(\vec{x}|v)f_1(v)|J|d\vec{x}dv = f(\vec{x};\theta)d\vec{x}$, and therefore

$$E(U) = \int Tf(\vec{x},\theta)d\vec{x} = E(T) = \theta\,,$$

since T is unbiased. The proof for discrete variables follows straight-forwardly from this. It is given below for our example.

The variance of U is less than (or equal to) that of T, because of the condition V fixed. If V and T belong to the bivariate normal sketched above, with means μ_V and μ_T, variances σ_V^2 and σ_T^2, and correlation coefficient ρ, then

$$U = E(T|V) = \mu_T - \rho\mu_V\sigma_T/\sigma_V + \rho\sigma_T V/\sigma_V ,$$

which is the line sketched in the figure. It is trivial to confirm that $E(U) = \mu_T = E(T)$ in this simple case. We can further easily verify that

$$\mathrm{Var}(U) = \rho^2\sigma_T^2 \leqslant \sigma_T^2 ,$$ so the conditioning gives us an improvement as expected.

In our example, since T = 1 only if there are no events in bin 1, and U is E(T|V), U is the expected frequency that a given number of events v, with Poisson parameter μ in each bin, will distribute themselves among n bins such that there are no events in the first bin. The p.d.f. for finding x_1 events in bin 1 for fixed v is a binomial, with the number of "successes" equal to the number of events in the first bin; i.e.,

$$f(x_1|v) = B(x_1;v,p)$$

$$= \frac{v!}{x_1!(v-x_1)!}\, p^{x_1}(1-p)^{v-x_1} ,$$

where p = the probability of one event going into bin 1; this is 1/n since all bins have the same p.d.f. Therefore,

$$U = E(T|v) = \sum_{x_1=0}^{\infty} Tf(x_1|v) = (1-p)^v = (1-\frac{1}{n})^v \text{ (inserting the value of T)} .$$

Therefore

$$E(U) = \sum_{v=0}^{\infty}(1-\frac{1}{n})^v\,\frac{e^{-n\mu}(n\mu)^v}{v!} = e^{-\mu}\sum_v e^{(n-1)\mu}\,\frac{[(n-1)\mu]^v}{v!} = e^{-\mu} .$$

Therefore, U is an unbiased estimator for $e^{-\mu}$, as planned.

$$V(U) = e^{-2\mu}[e^{\mu/n} - 1] \rightarrow 0 \text{ as } n \rightarrow \infty .$$

As an exercise, it can be shown that this estimator, $U = (1-\frac{1}{n})^v$, is asymptotically efficient (but not for any finite n). It can be shown also that it is unique, in being unbiased with this property.

Now we will state a general result for exponential families (and a few others).

Theorem: Let $x_1,\ldots,x_n$ be a sample from a p.d.f. $f(x,\theta)$, where θ may be multi-dimensional. Let f be a member of the exponential family, $f(x,\theta) = \exp\{A(\theta)C(x) + B(\theta) + K(x)\}$. Assume that we wish to estimate some function $g(\theta)$. Find any unbiased estimator for $g(\theta)$, let's say g_1, so $E(g_1) = g(\theta)$. Then $C(X)$ is a sufficient statistic for θ. If θ is a multidimensional vector, so is $C(x)$, with the same dimensions. Then the quantity $U = E(g_1|C)$ is independent of θ, and is therefore a statistic.

Further U has the properties that

a) $E(U) = g(\theta)$

b) U is the unique estimator of uniformly minimum variance.

All this is very well. How do we construct estimators?

Chapter 8. Constructing Estimators I — Substitution Methods

Frequency Substitution

This is useful in cases in which the population parameter to be estimated is a frequency or is a function of a frequency. In the method of frequency substitution, we simply replace unknown population frequencies by the experimentally observed frequencies.

Example — The Binomial

$$B(x; n,p) = \binom{n}{x} p^x(1 - p)^{n-x}$$

If we wish to know some function of p, e.g., g(p), we replace p by x/n, since E(x) = np. This works well for fairly large samples, where the law of large numbers says that the difference between E(x) and np is a very small fraction of the true value. The advantages are simplicity and that the estimator is usually consistent. The disadvantages are that the answer may be biased and may not have minimum variance. However, the bias may be small relative to other errors. We may be able to reduce this bias, or at least estimate its size, by a series expansion:

Bias Reduction

Suppose we begin with an unbiased estimator $\hat{\theta}$ for θ. We wish to estimate $g(\theta)$. We can try $\hat{g} = g(\hat{\theta})$, which will often have a slight bias. Then

$$g(\hat{\theta}) \simeq g(\theta) + (\hat{\theta} - \theta)g'(\theta) + \frac{1}{2}(\hat{\theta} - \theta)^2 g''(\theta) + \cdots$$

assuming g is twice differentiable, $g' = \dfrac{dg}{d\theta}$, etc.

Therefore $E[g(\hat{\theta})] \simeq g(\theta) + \frac{1}{2}V(\theta)g''(\theta) + \cdots$

since $E(\hat{\theta}) = \theta$ by assumption. Thus there is a bias in $g(\hat{\theta})$ equal to $\frac{1}{2}V(\theta)g''(\theta)$, to lowest order. We cannot calculate this term since we don't know θ. But if $\hat{\theta}$ is sufficiently close to θ, we can often satisfactorily estimate it by $\frac{1}{2}V(\hat{\theta})g''(\hat{\theta})$.

If we can calculate $V(\hat{\theta})$, and if $V(\hat{\theta})g''(\hat{\theta})$ is reasonably small, then we can make a bias-reducing correction to $g(\hat{\theta})$:

$$\hat{g}_1(\theta) = \hat{g} - \frac{1}{2}V(\hat{\theta})g''(\hat{\theta}) \ .$$

In the multivariate case, $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$,

$$\hat{g} = g(\hat{\theta}) \simeq g(\theta) + \sum_i (\hat{\theta}_i - \theta_i)\frac{\partial g}{\partial \theta_i}\Big|_{\theta_i} + \frac{1}{2}\sum_{i,j}(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)\frac{\partial^2 g}{\partial \theta_i \partial \theta_j}\Big|_{\theta_i,\theta_j} + \cdots ,$$

and

$$E(\hat{g}) \simeq g(\theta) + \frac{1}{2}\sum_{i,j} V_{ij}\frac{\partial^2 g}{\partial \theta_i \partial \theta_j}\Big|_{\theta_i,\theta_j} + \cdots ,$$

from which we deduce that

$$\hat{g}_1 = \hat{g} - \frac{1}{2}\sum_{i,j} V_{ij} \frac{\partial^2 g}{\partial\theta_i\partial\theta_j}\Big|_{\hat{\theta}_i,\hat{\theta}_j}$$

has reduced bias, provided the second term is not large and not rapidly-varying. If that proviso is not satisfied, it is not obvious that going to higher-order terms will help, since the problem may come from $\frac{1}{2}V(\hat{\theta})g''(\hat{\theta}) \neq \frac{1}{2}V(\theta)g''(\theta)$. In that case a more detailed analysis of the specific problem may be necessary, perhaps employing Monte Carlo techniques to help develop and test improved estimators.

If the correction is small, and its expectation value is small, it won't change the variance of $\hat{g}$ very much (probably). If it does affect it, it will often increase it; however, the MSE may be lower.

In the no-count Poisson example of the previous chapter, if we started with $\hat{\mu} = \overline{x}$ and therefore set $\hat{\theta} = e^{-\overline{x}}$, then $V(\hat{\mu}) = \mu/n$, and

$$E(e^{-\overline{x}}) \simeq e^{-\mu}\left(1 + \frac{\mu}{2n}\right) .$$

Therefore we can make a correction such that

$$\hat{\theta}_1 = e^{-\overline{x}}\left(1 - \frac{\overline{x}}{2n}\right) , \text{ assuming } \frac{\mu}{2n} \simeq \frac{\overline{x}}{2n} ;$$

further errors approach zero faster than 1/n as n → ∞.

Our best estimator was

$$\hat{\theta} = \left(1 - \frac{1}{n}\right)^{n\overline{x}} \quad \text{since } v = n\overline{x}$$

$$\simeq e^{-\overline{x}}\left(1 - \frac{\overline{x}}{2n}\right) \quad \begin{array}{l}\text{and we have recovered our}\\ \text{best estimator to order } 1/n^2.\end{array}$$

Note: In frequency substitution, the estimators are not always unique. We can choose from among them on the basis of MSE, lack of bias, or whatever.

Method of Moments

This is another substitution method. To estimate a function $q(\theta)$, we express $q(\theta)$ as a function of the population moments, then substitute sample moments for population moments. If all we want is the first moment, and that happens to be a frequency, this is same as above. Thus, we write

$q(\theta) = g(m_1, m_2, ..., m_n)$, where $m_j = E(X^j)$. This assumes that the necessary moments exist.

E.g., m_1 = mean

$m_2 = E(X^2) = V(X) + [E(X)]^2 = V(X) + m_1{}^2$,

Thus, to estimate the population variance, V(X), by the method of moments, we write

$\sigma^2 = m_2 - m_1^2$. Therefore we try

$$\hat{\sigma}^2 = \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2 = \frac{1}{n}(x_i - \overline{x})^2 .$$

We saw that for the normal $\frac{1}{n-1}\sum(x_i - \overline{x})^2$ was unbiased. Therefore $\hat{\sigma}^2$ above has a slight bias. For the Poisson, θ is both the population mean and the population variance. Therefore for the Poisson, we could use either

$\hat{\theta} = \hat{m}_1$ or $\hat{\theta} = \hat{m}_2 - \hat{m}_1^2$ as an estimator for the variance.

Therefore the method of moments does not necessarily provide a unique estimator.

Variance of the Moments

It can be shown in general that

$$V(\hat{m}_j) = \frac{1}{n}(m_{2j} - m_j^2) \text{ for moment } m_j ;$$

$$COV(\hat{m}_j,\hat{m}_k) = \frac{1}{n}(m_{j+k} - m_j m_k) \text{ , assuming these moments exist.}$$

In practice, moment m_{j+k}, etc. can be estimated by $\hat{m}_{j+k}$, etc. We must realize that high-order moments are seriously affected by the tails of the distribution; therefore if we require high order moments, we may get very bad answers using the method of moments, e.g., answers with very high variances.

By the law of large numbers, the average of anything tends to its expectation value, under the assumption that its variance is finite. Therefore moments estimators are at least consistent, i.e., they approach the right value as n → ∞.

Generalization of the Method of Moments

Instead of expressing $q(\theta)$ as a function of the simple moments $E(X^j)$ we could write the p.d.f. in terms of a set of some other functions (other than the moments) of the data points x_i. Then we could solve for each element of q (if q is a vector of more than one element) in terms of the expectation values of these functions. That is, don't use the moments m_j, which are the expectation values of polynomials, but the expectation values of some other functions of x which we choose by inspection of the p.d.f. We write the p.d.f. in terms of the unknowns θ and a set of functions of x_i. We calculate the expectation values of those functions in terms of the unknowns. Finally, we estimate those expectation values as the sample means of the functions, and solve for the unknowns.

Example:

Decay of a vector meson into two pseudo-scalar mesons. The data consist of sets of angles $(\cos\theta_i, \phi_i)$. We know the p.d.f is

$$f(\cos\theta,\phi) = \frac{3}{4\pi}[\frac{1}{2}(1 - \rho_{00}) + \frac{1}{2}(3\rho_{00} - 1)\cos^2\theta$$

$$- \rho_{1,-1}\sin^2\theta \cos2\phi - \sqrt{2}\, \mathrm{Re}\rho_{10} \sin2\theta \cos\phi] .$$

The unknowns are

$(\rho_{00}, \rho_{1,-1}, \mathrm{Re}\rho_{10})$.

Choose three functions so that we can estimate three unknowns. These functions might be

$g_1 = \cos^2\theta$

$g_2 = \sin^2\theta \cos 2\phi$

$g_3 = \sin 2\theta \cos \phi$

Then, e.g.,

$E(g_1) = \int g_1 f \, d\cos \theta \, d\phi$.

Since the trigonometric functions tend to be orthogonal, we are a bit lucky here. The expectations of our functions are trivially solved for the unknowns, viz,

$E(g_1) = 1/5 \; (1 + 2\rho_{00})$

$E(g_2) = - \; 4/5 \; \rho_{1,-1}$

$E(g_3) = - \; 4/5 \; \sqrt{2} \; \mathrm{Re}\rho_{00}$

Each expectation value depends on only one of the unknowns. Now we estimate *these* moments from the data:

$$1/5(1 + 2\rho_{00}) = \overline{g}_1 = \frac{1}{n} \sum\cos^2 \theta_i$$

$$- \; 4/5\rho_{1,-1} = \overline{g}_2 = \frac{1}{n}\sum\sin^2 \theta_i \cos2\phi_i$$

$$- \; 4/5 \; \sqrt{2} \; \mathrm{Re}\rho_{10} = \overline{g}_3 = \frac{1}{n} \sum \sin^2\theta_1 \cos \phi_i$$

and we solve for the ρ's. We can estimate the errors in them from $V(\overline{g}_i)$ using the techniques below and propagation of errors, which we will study in the next chapter.

To recapitulate, we attempt to construct functions g_k such that

$E(g_k) = q_k(\theta)$; $k = 1,...m$; where θ has m components we wish to estimate, and the q_k are the appropriate expectation values.

Then we estimate these functions from

$$\hat{q}_k = \frac{1}{n}\sum_i g_k(x_i) .$$

Having done so, we solve for $\hat{\theta}$. Obviously, we need the same number of g_k's as there are elements of θ.

Covariance of $\hat{q}$

We also estimate the covariance matrix for $\hat{q}$ the same way, from the sample variance:

$$V_{kk}(\hat{q}) \equiv V(\hat{q}_k) = (\frac{1}{n})^2 V[\sum_i g_k(x_i)] = \frac{1}{n} V[g_k(x)]\,,$$

since the x_i's are independent, identically distributed,

$$\simeq \frac{1}{n} \cdot \frac{1}{n-1} \sum_i [g_k(x_i) - \overline{g_k(x)}]^2\,.$$

The factor $\frac{1}{n-1}$ comes from the unbiased estimator of the sample variance, as we have studied earlier. The data sample is here regarded as a set of n×m points $g_k(x_i)$ rather than the x_i themselves.

The general term is

$$V_{kj}(\hat{q}) \simeq \frac{1}{n} \cdot \frac{1}{n-1} \sum_i (g_k(x_i) - \overline{g_k(x)})(g_j(x_i) - \overline{g_j(x)})\ .$$

We usually interpret $V(\hat{q}) = \sigma^2$ as the "square of the expected error" of $\hat{q}$. We sometimes write $q = \hat{q} \pm \Delta q$, with $\Delta q = \sigma$, an error symmetric about $\hat{q}$ as though $\hat{q}$ were normally distributed. If the error is known to be significantly skewed, as is often the case with maximum likelihood estimators (Chapter 10), this fact should be indicated in the error.

There are other expressions we could use to represent an error. For example, one could use the average of the absolute value of the error. The variance is conventional because (1) it is low order, therefore easy to calculate; (2) it is sufficient in the case of a Gaussian, because a Gaussian is *completely* described by μ and σ; (3) it is often consistent, because, for example, an exponential family will approach a Gaussian in the asymptotic limit, thanks to the C.L.T.; (4) it can be easily converted to a probability in the Gaussian (C.L.T.) limit. We will, however, see a case in the next chapter in which this estimate of the error is very flawed.

We now need to know how to convert our covariance matrix for $\hat{q}$ into a covariance matrix for $\hat{\theta}$.

Chapter 9. Propagation of Errors

General

This is a technique needed for substitution methods of estimation, wherein one estimates some variable $\hat{q}$ and its variance or MSE, but one then substitutes $\hat{q}$ into some function $\theta(q)$, obtaining $\hat{\theta} = \theta(\hat{q})$. The true value of q is q_t, and the corresponding value of θ, the desired object of our investigation, is $\theta(q_t)$. $\hat{q}$ is distributed about q_t according to some p.d.f. We wish to approximately evaluate the variance of $\hat{\theta}$ from the estimated variance of $\hat{q}$. We will assume $\hat{q}$ is an unbiased estimator of q.

Expand $\hat{\theta}$ about q_t:

$$\hat{\theta} = \theta(\hat{q}) \simeq \theta(q_t) + \left.\frac{\partial\theta}{\partial q}\right|_{q=q_t} (\hat{q} - q_t) + \cdots$$

Therefore

$$E(\hat{\theta}) \simeq \theta(q_t) + \text{second and higher order terms [since } E(\hat{q} - q_t) = 0] \ .$$

The removal of the first-order term justifies using $\hat{\theta}(q) = \theta(\hat{q})$.

$$\text{Now, } [\hat{\theta} - E(\hat{\theta})]^2 \simeq [\theta(\hat{q}) - \theta(q_t)]^2 = \left[\left.\frac{\partial\theta}{\partial q}\right|_{\hat{q}}\right]^2 (q_t - \hat{q})^2 + \cdots ,$$

and hence we can estimate the variance of $\hat{\theta}$:

$$E\{[\hat{\theta} - E(\hat{\theta})]^2\} \equiv V(\hat{\theta}) = \left[\left.\frac{\partial\theta}{\partial q}\right|_{\hat{q}}\right]^2 V(\hat{q}) + \cdots .$$

This is the technique of propagation errors.

Note that $V(\hat{\theta})$ is not the MSE unless the bias of $\hat{\theta}$ is zero. In the technique of bias reduction we saw a little earlier (Chapter 8), we had to go to second order in the Taylor series. Here, we only need to go to first order, for most applications. This technique works well only when second and higher order terms are small.

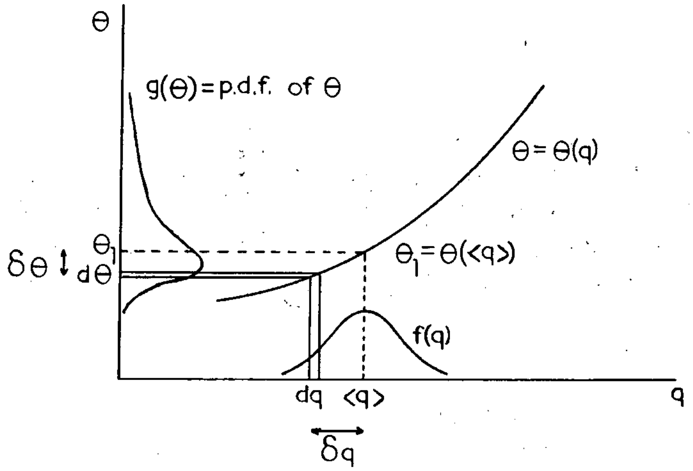
Example

$$\theta(q) = A + Bq \ .$$

Propagation of errors derives $V(\hat{\theta}) = B^2V(\hat{q})$, which is the exact result we have derived before. Any bias depends upon the bias of $\hat{q}$ and the correctness of A and B for the problem at hand. Since θ is a completely linear function of q, all higher order terms in the propagation of errors drop out, which is why we achieve the exact result.

The general case bears a close resemblance to the change of variables problems we have already discussed. Here, however, it is assumed that it is not possible or not convenient to do a full change of variables analysis; nevertheless we require a reasonable approximation to the true variance.

Geometrically,



In the sketch, we write $\langle q \rangle$ as the expectation value of $\hat{q}$. The shape of $g(\hat{\theta})$, the p.d.f. for $\hat{\theta}$, is altered from $f(\hat{q})$ by $d\theta/dq$, the Jacobean. In regions where $d\theta < dq$, the probability piles up faster for θ than for q, so the p.d.f. for θ is higher than for q. Thus, in the sketch, the peak in $g(\hat{\theta})$ occurs below $\theta_1 = \theta(\langle q \rangle)$.

If $f(\hat{q})$ is normal, $g(\hat{\theta})$ is not normal unless $\theta(q)$ is linear. This is the source of possible biases in $\hat{\theta}$. If $f(\hat{q})$ is symmetric about $\langle q \rangle$ (let's say $\langle q \rangle$ happens to be the true mean), then $\frac{1}{2}$ of the probability lies above $\hat{\theta} = \theta_1$, $\frac{1}{2}$ below, i.e., θ_1 is at the median. In the above figure, the mean of $g(\hat{\theta})$ is probably above θ_1. If $\theta(q)$ curves rapidly, higher terms are important, the bias may be large, and errors in $\hat{\theta}$ may be quite asymmetric. This is an important case.

The multidimensional version of the above is as follows:

$$\vec{q} = (q_1, \ldots, q_n).$$

(A frequent case involves the multivariate normal, where $\vec{q}$ has n × n covariance matrix V.) The dimensionality of θ may be smaller than that of q; if it is larger, we don't have adequate data. Now

$$\hat{\theta}_i = \theta_i(\hat{q}) \simeq \theta_i(\vec{q}_t) + \frac{\partial \theta_i}{\partial q_1}(\hat{q}_1 - q_{1t}) + \frac{\partial \theta_i}{\partial q_2}(\hat{q}_2 - q_{2t})$$

$$+ \ldots + \frac{\partial \theta_i}{\partial q_n}(\hat{q}_n - q_{nt}) + \text{second and higher order terms.}$$

As written here, θ_i is the i^th component of a multi-dimensional vector. Therefore

$$E(\hat{\theta}) \simeq \theta(q_t) \quad \text{(dropping the subscript on theta), and}$$

$$[\hat{\theta} - \theta(q_t)]^2 \simeq \sum_{i,j} \left(\frac{\partial \theta}{\partial q_i} \right) \left(\frac{\partial \theta}{\partial q_i} \right) (q_{it} - \hat{q}_i)(q_{jt} - \hat{q}_j)$$

Therefore

$$V(\hat{\theta}) \simeq \vec{D}^T(\theta) V(\hat{q})\, \vec{D}(\theta)\,,$$

$$\text{where } \vec{D}(\theta) = \begin{pmatrix} \frac{\partial\theta}{\partial q_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial\theta}{\partial q_n} \end{pmatrix}_{q=\hat{q}}$$

This is a congruence transformation [Basilevsky (1983), Theorem 4.6] which carries the positive definite matrix $V(\hat{q})$ into another positive-definite matrix $V(\hat{\theta})$. If our change of variables is linear, it will carry a multivariate normal into another multivariate normal; otherwise not. However, if $\vec{\theta}$ is an arbitrary coordinate vector in $\hat{\theta}$-space, the quadratic form $y = \vec{\theta}^T V(\hat{\theta})\vec{\theta}$ is still positive and the contour y = constant is still an ellipse. We still write the covariance matrix V as the error matrix even though q may not be a multivariate normal. If θ is multi-dimensional, we can do this for each element separately.

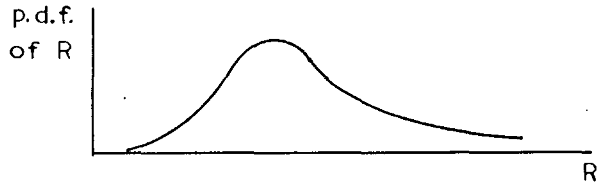
Let's look at a bad case for propagation of errors: F. James (1983) analyzed the data of an experiment measuring the mass of the neutrino. The variable θ is represented by R:

$$R = \frac{a}{\frac{d}{K^2e}(b - c) - 2(1 - \frac{K^2d}{Ke})a} \, ,$$

where a,b,c,d, and e are measured and K is fixed. If $R < 0.420$, we must conclude that the neutrino has a non-zero mass, otherwise it is consistent with having zero mass. They measured $R = 0.165$. Calculating the errors with propagation of errors, the experiment arrived at $\Delta R = 0.073$, a symmetric error of course. Therefore 0.42 is 3 standard deviations away from the measured value, with about 1 chance in 1000 that $R \geqslant 0.42$. Therefore, there appears to be strong evidence for a non-zero mass.

But R is highly non-linear in the measured quantities. Some of the variables appear twice and therefore there may be large correlations, even if each of the a,b,c,d, and e are independent. Also, some of the errors are large ($\sim$30%). In particular, and perhaps most troublesome, some of the terms in the denominator have large errors. James set up a Monte Carlo calculation, which assumed independent normals for a,b,c,d, and e with σ's as quoted by the experimenters. He found that 1-1/2% of the time (incorrectly quoted at 4% in the James reference) R came out $\geqslant 0.42$. This is a quite reasonable level of probability. There's a big difference between 0.1% and 1-1/2%. A correct calculation would take account of correlations and possible deviations from the normal for the input variables. However, all that is usually quite simple to do in a Monte Carlo calculation, and is a tremendous improvement over propagation of errors.

The point is, the p.d.f. of R comes out with a long asymmetric tail just where it would be significant for our conclusions (the high side); therefore neglect of the asymmetry has serious consequences, even if there is no net bias.



A general *rule of thumb:* if your transformation depends on $1/q_i$ in some way, watch out! It may well be both biased and improperly handled by propagation of errors (unless the errors on the denominator are small). This case may be better handled by quoting an asymmetric "confidence interval" for R, rather than symmetric errors.

We can go to a higher order in the propagation of errors expansion, which is sometimes useful. But if $g(\hat{\theta})$ is skewed, we've got to be careful lest all we compute is corrections to the covariance matrix which we still interpret as giving symmetric errors. There is a nice discussion of higher-order corrections to propagation of errors in Meyer (1975).

This is more than an example of propagation errors, it is also an example of an hypothesis test. In this particular test, we ask for the probability that R is equal to a certain value "or worse," i.e., or greater. Hypothesis testing is covered in detail in the references at the back.

The principle use for an error estimate in general may be for some such hypothesis test (which may occur in the future). For example, is this theory (or maybe that one) consistent with the data? Another form of this same problem: given such and such a theory, our measurement is used as an input parameter, from which a prediction for some other quantity with error will be derived. This derived quantity is then to be compared with an experiment.

Discussion of Errors

Error estimates provide a feel for the range of values in which θ could be found, with "reasonable" probability, in light of the data. They are therefore probabilistic statements, which are usually interpreted on the basis of a normal p.d.f. That is, the interval is assumed to cover the true value of θ with a probability determined by a normal of mean value $\hat{\theta}$ and error as specified.

Often we are in a situation where we want to know whether or not some particular model, which predicts a certain value θ_1 for θ, is in agreement with the data. Many people phrase this in the form of a question about the consistency of "θ_1 or worse (i.e., farther away from the estimate $\hat{\theta}$)," because if we were to accept one of these worse values of θ we would surely accept θ_1. If $\hat{\theta}$ is Gaussian with an error which does not depend upon θ, we can perform a simple integration over the region in question (giving us an error function). We can also extract the term $(\hat{\theta} - \theta_1)^2/\sigma^2$ out of the Gaussian and look this up in a table of χ^2 cumulative distribution functions. These are available in most statistics books. Note, however, that the χ^2 takes the same values for θ_1 greater than $\hat{\theta}$ as for less than; we therefore do not distinguish one side from the other with the χ^2, except that we can recognize the symmetry of the Gaussian and divide the c.d.f. by two if we are only interested in one side.

If σ^2 is unknown and must be estimated from the data, we naturally want to increase the size of the region in θ which we consider to be in "good agreement" with the data. This must be done to allow for the fact that we may have underestimated σ^2. In this case, one no longer assumes the distribution of the error about $\hat{\theta}$ to be Gaussian; the "student's t distribution," similar to a Gaussian but with long tails, is used instead (if σ^2 is assumed not to vary with θ). This case is discussed in Chapter 11.

It may also, very frequently, be true that σ^2 will vary with θ. For example, if a bin of a histogram is filled by a Poisson process, the variance equals the mean. To assess the agreement between the observed number of counts and some predicted number, we need to recognize that, if the predicted number is to be assumed to represent the expectation value of the Poisson, it also represents the variance expected in an observation. This we do in such procedures as least-squares fitting to a histogram. Otherwise, even in the limit of large numbers in which the normal approximation to the Poisson is adequate, if we happen to underestimate θ_{true} we will quote errors which are too small, and conversely if we overestimate θ_{true}.

That is, we assign errors on the basis of the assumed θ_{true} rather than the observed data.

In more complex cases, it may be impossible to do this. In the example studied by James, there is no way to choose from among the infinite number of sets of values of a,...,e which could give R = 0.42. We also are not told how $\sigma_a^2, \ldots, \sigma_e^2$ might be affected by changes in the associated parameters, although the original experimenters might be able to estimate that.

The lesson to be learned is that it is wise to take error estimates, and their implied probabilities, very cautiously. Most physicists would probably not regard probabilities like 0.1% as being convincing evidence against a model. It is suggestive, but needs confirmation.

In addition to the problems we have touched upon, there are possible systematic errors. These may be loosely defined as reproducible but unknown inaccuracies in the apparatus or procedures which would lead to a bias, even if the experiment were capable of being repeated a large number of times so that random statistical errors were reduced to any desired level. Systematic errors are characteristically extremely difficult to estimate. They do not have a distribution except in the sense that similarly prepared experiments may have different values for the same type of error. Systematic errors may also vary through an experiment in a (possibly unknown) way, depending upon operating conditions or the age of components. For these reasons, many experimenters quote their statistical and systematic errors separately, for example, as

$$A = B \pm C(\text{stat.}) \pm D(\text{syst.}).$$

Since systematic errors vary in character with the experiment, there is no generally-accepted technique for calculating probabilities from them. Some people just add the systematic errors in quadrature with the statistical errors, but one must recognize that this (and any other technique) is probably wrong and therefore be very conservative in interpreting the results if systematic errors are large.

Chapter 10. Constructing Estimators II — Maximum Likelihood

This technique is very generally used because it is often the simplest method — perhaps the only approach in complex cases — and because it provides estimates which have certain desirable properties. The basic idea is that we suppose that a measurement x is the result of a sampling of some p.d.f. which depends upon a set of parameters $\vec{\theta} = (\theta_1, \ldots, \theta_k)$, i.e., $f(x;\vec{\theta})$. Let us for the beginning assume θ is only one-dimensional. Now we argue that nature has provided us with our sampling, which depends only upon fixed but unknown θ. The probable range of θ is assumed to be restricted by the assumption that if our specific sample were unlikely to have arisen from some value of $\theta = \theta_1$, then conversely it is unlikely that θ_1 is the true value. We pick as our best guess for θ that value such that the joint probability of our sample is maximized.

That is, if $\vec{x} = (x_1,\ldots,x_n)$ is our sample, the joint p.d.f. is

$$\prod_{i=1}^{n} f_i(x_i;\theta)$$

if the x_i's are independent, or, more generally, $g(\vec{x};\theta)$ if they are not independent. This normally means "the probability of $\vec{x}$ given fixed θ." We now reverse the interpretation to fix $\vec{x}$ at the measured values, and allow θ to vary. The real θ doesn't vary, of course, but our estimate of it does.

Note: θ is allowed to take on a continuous set of values, or may just refer to one of a finite set of models which we are testing.

We construct a likelihood function

$$\begin{aligned}\mathscr{L}(\theta) &= \prod_i f_i(x_i,\theta) \quad (x_i \text{ independent}) \\ &= g(\vec{x},\theta) \quad (x_i \text{ not independent}) \; .\end{aligned}$$

The likelihood is equal to (or proportional to, see below) the joint p.d.f. of the x_i, but it is evaluated at fixed x_i, the measured values of the r.v.'s, and varies as a function of θ. It is *not* a p.d.f. for θ; if it were, we would take its expectation value as $\hat{\theta}$. Instead, we evaluate $\hat{\theta}$ from:

The Principle of Maximum Likelihood

The best explanation for a set of data is provided by that value of θ which maximizes the likelihood function.

This is not capable of proof without a definition of "best". We will show that the M.L. value for θ has a number of desirable properties; however, it is often slightly biased. The Principle of Maximum Likelihood should be treated as a heuristic principle. That is, we use it where it works well, but we should be prepared to use another technique (where possible) or to modify the M.L. estimator in the event of an unacceptable bias or other problem.

We will throughout only be interested in relative values of the likelihood, such as its maximum. Therefore the likelihood can always be multiplied by an arbitrary constant which is independent of θ, but which is allowed to depend upon x.

Example

A certain experiment has just three possible outcomes, Z_1, Z_2, and Z_3. There are three different models or theories to explain the physical process which determines the outcome. These models are labeled θ_1, θ_2, and θ_3. For each model the probability of observing each of the outcomes is listed in this table:

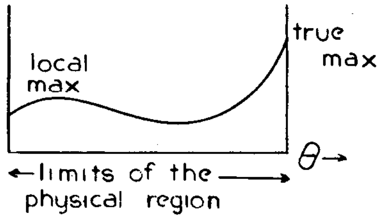
Possible		Probabilities			
Outcomes		$P(Z\|\theta_1)$	$P(Z\|\theta_2)$	$P(Z\|\theta_3)$	
		θ_1	θ_2	θ_3	← competing models
Z_1	$\mathscr{L}(\theta\|Z_1) \rightarrow$	.4	.6	.2	
Z_2	$\mathscr{L}(\theta\|Z_2) \rightarrow$	.2	.3	.1	
Z_3	$\mathscr{L}(\theta\|Z_3) \rightarrow$	.4	.1	.7	
		1.0	1.0	1.0	← the sum of the probabilities

The likelihood function starts with a certain fixed outcome and varies with θ. The M.L. principle asserts that we should be able to work back from a measured outcome and choose as our "best" θ that value which maximizes $\mathscr{L}$. If we get outcome Z_2, for example, we should argue that θ_2 is our maximum likelihood estimate. In no case would we choose θ_1. Notice that the rows don't add up to 1, i.e., the $\mathscr{L}$'s are not p.d.f.'s. Notice that $\mathscr{L}(\theta|Z_1) = 2 \times \mathscr{L}(\theta|Z_2)$. We will therefore draw exactly the same conclusion about θ from Z_1 and Z_2. That is, the likelihood technique cannot distinguish outcomes with proportional likelihood functions.

In M.L. calculations, we usually work from $\ell \equiv \ell n\mathscr{L}$, rather than $\mathscr{L}$ directly, because it is usually simpler to work with. It converts products of p.d.f.'s into sums over the logarithms of p.d.f.'s, and it removes exponentials, e.g., in Gaussians. The maximum of ℓ is at the same θ as the maximum of $\mathscr{L}$:

$$\frac{\partial \ell n\mathscr{L}}{\partial\theta} = \frac{1}{\mathscr{L}}\frac{\partial\mathscr{L}}{\partial\theta} = 0$$

gives us $\hat{\theta}$. This is called the likelihood condition. This finds an extremum which may be a minimum; therefore it is important to check. There may also be > 1 local maximum; we usually take the highest, if we can find it. But this leads to ambiguities, i.e., multiple solutions. Multiple solutions usually go away in the asymptotic region, i.e., for large amounts of data. We might even have saddle points in special cases: $\mathscr{L}$ as a function of continuous θ_1 and θ_2 might attain a maximum for θ_1 but at a minimum for θ_2, for example. It can also happen that the maximum occurs at a boundary, e.g.,



In this case, the likelihood equation does not suffice to find the maximum

We might hope that as n → ∞, a unique, clear maximum in the physical region asserts itself; this is usually, but not always, the case.

Cases in which we don't want to use $\ell n \mathscr{L}$ are encountered in complex problems. For example, a subroutine full of IF statements and complex equations or Monte Carlo estimates is used to return a normalized p.d.f. for $\vec{x}$ given θ. This p.d.f. is used as a likelihood. The normalization is only needed to the extent that it has a θ dependence, otherwise it just goes into the arbitrary constant by which we can multiply $\mathscr{L}$.

Example: Normal $N(x_i; \mu_i, \sigma_i)$

$$\mathscr{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}} \quad ; \ \theta_i = (\mu_i, \sigma_i) \text{ , all } x_i\text{'s independent.}$$

Notice that we construct $\mathscr{L}$ out of p.d.f.'s, not probabilities, except in the discrete variable case. We could convert these density functions to probabilities by multiplying by the product of dx_i's, but since that small volume element in $\vec{x}$ space doesn't depend upon θ, we can ignore that complication. Continuing,

$$\ell = \sum_i \left\{ -\frac{1}{2} \ell n\,(2\pi) - \ell n\,\sigma_i - \frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right\} .$$

The ignorable multiplicative constants in $\mathscr{L}$ become ignorable additive constants in $\ell = \ell n \mathscr{L}$.

We ignore here the case in which all observations are equal, which gives problems because finding the maximum requires that $\sigma_i \to 0$ and $\mathscr{L} \to \infty$ (if we are estimating σ_i).

Solving the likelihood equations:

$$\frac{\partial \ell}{\partial \mu_i} = +\frac{2}{2} \frac{(x_i - \mu_i)}{\sigma_i^2} = 0$$

$$\frac{\partial \ell}{\partial \sigma_i} = -\frac{1}{\sigma_i} + \frac{2}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^3} = 0 .$$

This gives us 2n equations in 2n unknowns. An important case occurs if we assume $\mu_i = \mu$ for all i, and that all the σ_i's are known. Then

$$\frac{\partial \ell}{\partial \mu} = \Sigma \left(\frac{x_i - \mu_i}{\sigma_i^2} \right) = \Sigma \frac{x_i}{\sigma_i^2} - \Sigma \frac{\mu}{\sigma_i^2}$$

The M.L. principle asserts that this = 0 if $\mu = \hat{\mu}$:

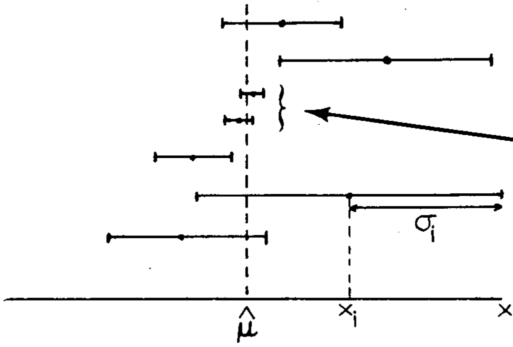
$$\Sigma \frac{x_i}{\sigma_i^2} = \hat{\mu} \Sigma \frac{1}{\sigma_i^2} \text{ , and therefore}$$

$$\hat{\mu} = \frac{\Sigma x_i/\sigma_i^2}{\Sigma \frac{1}{\sigma_i^2}} \text{ . This is a weighted average.}$$

Check for bias:

$$E(\hat{\mu}) = \frac{\sum \frac{E(x_i)}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}} = \frac{\sum \frac{\mu}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}} = \mu \text{ . OK.}$$

For example:



mostly determined by these 2 data.

Check the variance:

$$V(\hat{\mu}) = E(\hat{\mu}^2) - \mu^2 = \left(\frac{1}{\sum \frac{1}{\sigma_i^2}}\right)^2 E\left[\left(\sum \frac{x_i}{\sigma_i^2}\right)^2\right] - \mu^2 \,;$$

all the x_i's are independent, therefore

$$\begin{aligned} E(x_i x_j) &= \mu\mu; && i \neq j \\ &= \sigma^2 + \mu^2; && i = j\,. \end{aligned}$$

Therefore,

$$V(\hat{\mu}) = \left(\frac{1}{\sum \frac{1}{\sigma_i^2}}\right)^2 \left[\sum_i \frac{\sigma_i^2 + \mu^2}{\sigma_i^4} + \sum_{i \neq j} \frac{\mu^2}{\sigma_i^2 \sigma_j^2}\right] - \mu^2$$

$$= \left(\frac{1}{\sum \frac{1}{\sigma_i^2}}\right)^2 \left[\sum \frac{1}{\sigma_i^2} + \mu^2 \sum \frac{1}{\sigma_i^4} + \mu^2 \sum_{i \neq j} \frac{1}{\sigma_i^2 \sigma_j^2}\right] - \mu^2$$

$$= \left(\frac{1}{\sum \frac{1}{\sigma_i^2}}\right) + \mu^2 \left[\frac{\sum \frac{1}{\sigma_i^4} + \sum_{i \neq j} \frac{1}{\sigma_i^2 \sigma_j^2} - \left(\sum \frac{1}{\sigma_i^2}\right)^2}{\left(\sum \frac{1}{\sigma_i^2}\right)^2}\right]$$

$$= \frac{1}{\sum 1/\sigma_i^2} \; .$$

The R.C.F. Inequality (Chapter 6) says

$$V(\hat{\mu}) \geqslant \frac{1}{I(\mu)} \; ;$$

$$I(\mu) = -E\left[\frac{\partial^2}{\partial\mu^2}\sum \ell n\, f(x_i,\mu)\right]$$

$$= -E\left[\frac{\partial^2}{\partial\mu^2}\ell\right]$$

We already know $\frac{\partial\ell}{\partial\mu}$, which equals zero at $\mu = \hat{\mu}$ by the likelihood principle. Therefore

$$I(\mu) = -E\left[\frac{\partial}{\partial\mu}\left(\sum\frac{x_i}{\sigma_i^2} - \sum\frac{\mu}{\sigma_i^2}\right)\right]$$

$$= -E\left[-\sum\frac{1}{\sigma_i^2}\right] = \sum\frac{1}{\sigma_i^2} \; .$$

Therefore the M.L. solution $\hat{\mu}$ is completely efficient, and we have found the best possible estimator, assuming the Gaussian model is correct. This is no accident. The method of maximum likelihood will find the efficient estimator if one exists. This will be demonstrated now.

Let us assume that there exists an unbiased, efficient estimator T(x). Then we have proved earlier (Chapter 6) that T(x) is linearly related to $S(\vec{x},\theta) = \frac{\partial}{\partial\theta}\ell n\, f(\vec{x},\theta)$, the *score function.* This relationship must be of the form

$$S(\vec{x},\theta) = C(\theta)T(\vec{x}) + D(\theta),$$ where C and D are not functions of $\vec{x}$, and T is not a function of θ.

The likelihood condition gives us

$$\frac{\partial}{\partial\theta}\ell n\, f(\vec{x},\theta)\Big|_{\theta=\hat{\theta}} = S(\vec{x},\hat{\theta}) = 0 \; .$$

(Here, we obviously restrict ourselves to cases in which the maximum of $\mathscr{L}$ is not at a boundary.) This alone tells us only that $\hat{\theta}$ is such that $T(\vec{x}) = -D(\hat{\theta})/C(\hat{\theta})$. We also know that the score function has the property that

$$E[S(\vec{x},\theta)] = E\left[\frac{\partial}{\partial\theta}\ell n\, f(\vec{x},\theta)\right] = \frac{\partial}{\partial\theta}\int f\, dx = \frac{\partial}{\partial\theta}(1) = 0 \, ,$$

under certain very general conditions on f. This holds for any θ in the allowed range; it doesn't have to be the true θ. In other words, this holds for all members of the family of the p.d.f. In particular, it is true at $\theta = \hat{\theta}$, the M.L. estimator:

$$E[S(\vec{x},\hat{\theta})] \equiv E[S|\hat{\theta}] = C(\hat{\theta})E[T|\hat{\theta}] + D(\hat{\theta}) = 0 .$$

Hence, $E[T(\vec{x})|\hat{\theta}] = T(\vec{x})$. (This requires proving, because, even though T doesn't depend upon θ, the expectation does.) We assume that a solution $\hat{\theta}$ satisfying this condition exists in the allowed range. At first glance, the condition seems paradoxical, because

$$E[T(\vec{x})|\hat{\theta}] = \int T(\vec{y})\, f(\vec{y},\hat{\theta})d\vec{y} \qquad (\vec{y} \text{ a dummy variable})$$

doesn't appear to be a function of $\vec{x}$. The paradox is resolved if we remember that $\hat{\theta}$ is chosen by the condition $T(\vec{x}) = -D(\hat{\theta})/C(\hat{\theta})$; i.e., $\hat{\theta}$ itself depends upon the data $\vec{x}$ (after all, $\hat{\theta}$ is a statistic).

Now we do the final step:

By construction, $\int T(\vec{y})\, f(\vec{y},\theta)d\vec{y} = \theta$, since T is an unbiased estimator for the family $f(\vec{y},\theta)$ for any θ. Since this is just a mathematical statement not dependent upon the actual value of θ, it must be true for $\theta = \hat{\theta}$, the M.L. solution. Therefore

$$E[T(\vec{x})|\hat{\theta}] = T(\vec{x}) = \hat{\theta} ,$$

and the Likelihood Principle finds the unbiased, efficient estimator $T(\vec{x})$.

A few remarks:

(a) Evidently, $\hat{\theta} = -D(\hat{\theta})/C(\hat{\theta})$; therefore $D(\hat{\theta}) = -\hat{\theta}\,C(\hat{\theta})$ and $D(\theta) = -\theta\,C(\theta)$ (this is true because $\hat{\theta} = -D(\hat{\theta})/C(\hat{\theta})$ must hold for any $\hat{\theta}$ which may be selected by $T(x) = \hat{\theta}$); hence $S(\vec{x},\theta) = C(\theta)[\hat{\theta} - \theta]$. This is a general result for efficient, unbiased, estimators. If the estimator is biased but efficient, $S(\vec{x},\theta) = C(\theta)[\hat{\theta} - b(\theta) - \theta]$, where $b(\theta)$ is the bias: $E(\hat{\theta}) = \theta + b(\theta)$.

(b) The variance of $\hat{\theta}$ may be obtained from the R.C.F. Inequality, which becomes an equality here by assumption:

$$V(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{1}{E[S^2]} = \frac{-1}{E[\frac{\partial S}{\partial \theta}]} .$$

From (a) we obtain:

$$\frac{\partial S}{\partial \theta} = C'(\theta)\hat{\theta} + D'(\theta) = C'(\theta)\hat{\theta} - C(\theta) - \theta C'(\theta) ,$$

where the prime denotes differentiation. In this calculation, θ is the true value of the parameter, assumed known. This derivative expression is a linear function of $\hat{\theta}$, and $E(\hat{\theta}) = \theta$. Therefore,

$$E\left[\frac{\partial S}{\partial \theta}\right] = -C(\theta) \quad \text{(since E is a linear operator);}$$

therefore $V(\hat{\theta}) = 1/C(\theta) > 0$. We will later see that we will estimate $V(\hat{\theta})$ by $1/C(\hat{\theta})$ in experimental situations.

(c) At any value of θ, not necessarily the true value of the parameter,

$$\frac{\partial^2 \ell n \mathscr{L}}{\partial \theta^2} = \frac{\partial S}{\partial \theta} = C'(\theta)(\hat{\theta} - \theta) - C(\theta)$$

$$= - C(\hat{\theta}) \text{ at } \theta = \hat{\theta} .$$

From (b), this must be negative, as long as $\hat{\theta}$ is reasonably close to θ_{true}. Therefore the solution $\hat{\theta}$ of the likelihood equation $S = 0$ is a maximum as desired.

(d) From the form of the score function in (a), the M.L. solution $\hat{\theta}$ is unique, since (b) assures us $C(\theta) \neq 0$.

(e) all of the above assume the data $\vec{x}$ are relevant to estimating θ. This assumption is (almost) guaranteed by the premise, that there exists an estimator $T(\vec{x})$ unbiased and efficient.

In the Gaussian case again, suppose both μ and σ are unknown, but $\sigma_i = \sigma$ for all i.

$$\frac{\partial \ell}{\partial \mu} = \sum_i \left(\frac{x_i - \mu}{\sigma} \right) = 0$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_i \left(-\frac{1}{\sigma} + \frac{x_i - \mu}{\sigma^3} \right) = 0$$

From the first one, we get

$\hat{\mu} = \frac{1}{n} \sum_i x_i$. Substituting in the second one, we get

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \overline{x})^2$$

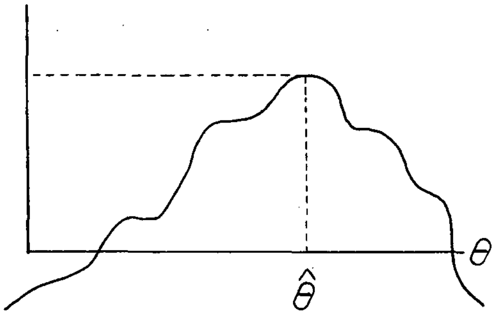
this is a biased estimate for σ. An unbiased estimate is $\frac{1}{n-1} \sum (x_i - \overline{x})^2$, as we have shown earlier (Chapter 5). Note: If μ were known, not having to be estimated, we would get, from the likelihood condition,

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \mu)^2 .$$

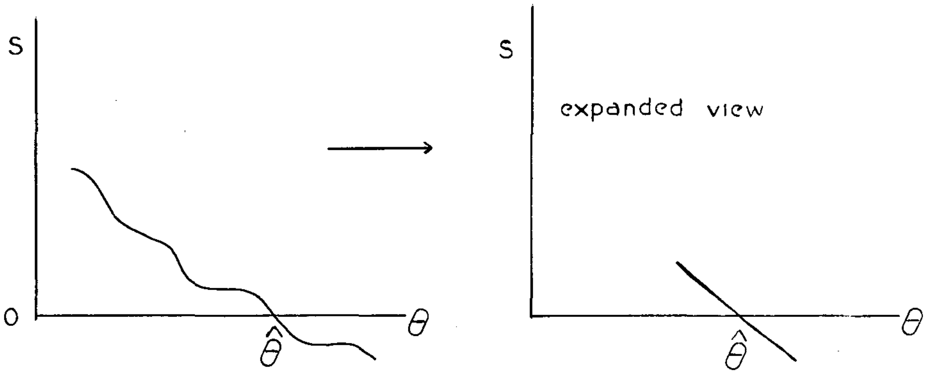
This is almost the same thing, but now this is unbiased, because μ is known.

Asymptotic Properties

Maximum Likelihood estimates are, in general, neither unbiased nor efficient. However, they are (usually) both *asymptotically,* again because of the connection of the M.L. estimator to the score: Let us look at the form of $\ell = \ell n \mathscr{L}$:



$$\text{Score } S(x,\theta) = \frac{\partial}{\partial\theta}\sum_i \ell n\, f(x_i,\theta) \simeq S(x,\hat\theta) + \frac{\partial S}{\partial\theta}\Big|_{\hat\theta}(\theta-\hat\theta) + \cdots$$



(Note: $\frac{\partial^2 \ell n\, f}{\partial\theta^2} = \frac{\partial S}{\partial\theta} < 0$ near the solution, because the solution is a maximum.) Here, again, we make our calculation as if we knew θ. We want to find results good for all θ, if possible. Under suitable conditions, the higher order terms in this expansion of the score → 0 as n → ∞. Therefore,

$$S(\vec{x},\theta) \simeq \frac{\partial S}{\partial\theta}\Big|_{\hat\theta}(\theta-\hat\theta) \quad \text{(since } S(\vec{x},\hat\theta) = 0 \text{ by choice of } \hat\theta\text{)}$$

$$= \frac{\partial^2}{\partial\theta^2}\sum_i \ell n\, f(x_i,\theta)\Big|_{\hat\theta}(\theta-\hat\theta)$$

$$\simeq E\left[\left(\frac{\partial^2}{\partial\theta^2}\sum \ell n\, f\Big|_{\hat\theta}\ \hat\theta\right)\right](\theta-\hat\theta)\,,$$

where the approximation in the last step is to use the expectation value for the sample mean (i.e., the sample mean → expectation value as n → ∞, provided only that the variance is finite).

$$= -\, I(\hat\theta)(\theta - \hat\theta) \simeq -\, I(\theta)(\theta - \hat\theta)\,,$$

where the last step is exact in the limit we are considering, of a linear score function. This result holds because of the law of large numbers. Therefore in the asymptotic limit, if the law of large numbers applies, the score is a linear function of $\hat\theta$. Hence, in this limit, $\hat\theta$ is efficient, unbiased, and therefore consistent. Further, since

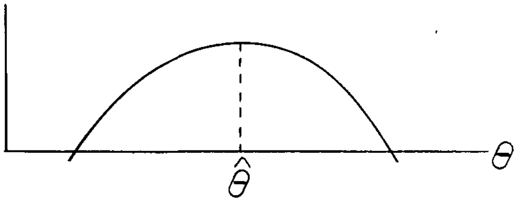
$$\frac{\partial}{\partial\theta}\ell n\mathscr{L} = S(\vec{x},\theta) \simeq I(\hat\theta)(\hat\theta - \theta)\,, \text{ we can write}$$

$$\ell n\mathscr{L} \simeq -\frac{I(\hat\theta)}{2}(\hat\theta - \theta)^2 + \ell n\, k, \text{ k a constant, and}$$

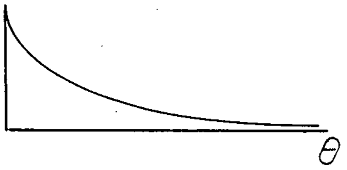
$$\mathscr{L}(\theta) \simeq k \exp(-(\hat\theta - \theta)^2\, I(\hat\theta)/2)\,.$$

That is, $\mathscr{L}$ is a Gaussian whose peak is at $\hat\theta$ with variance $1/I(\hat\theta)$. The latter occurs because the estimator is efficient. Notice that we are talking about the experimentally observed likelihood, where $\hat\theta$ is fixed by the data and θ is a variable; in spite of its form, $\mathscr{L}$ is not a p.d.f. However, we estimate the true mean with $\hat\theta$, and likewise we will now estimate the true variance with $1/I(\hat\theta)$. In the limit of a truly linear score, $I(\hat\theta) = I(\theta_{true})$, but we often assume this is approximately true even if the score is not linear. More on this later.

If $\mathscr{L}$ is Gaussian, $\ell = \ell n\mathscr{L}$ is parabolic:



Therefore, under suitable conditions, $\hat\theta$ is unbiased and of minimum variance, asymptotically. It should be obvious that one example of "unsuitable" conditions is the case in which the true maximum occurs at a boundary:



In general, for the conditions to be suitable requires the following:

— that $\theta \in \Omega$ which is finite dimensional and closed, and θ is not on the boundary of Ω.

— that the p.d.f.'s defined by different values of θ are distinct.

— that the first three derivatives of $\ell = \ell n\mathscr{L}$ exist in the neighborhood of the solution.

— that the information is finite and positive definite.

In case θ is a vector, we replace $1/I(\theta)$ by $I^{-1}(\theta)$ as the covariance matrix.

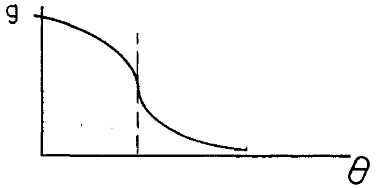
This property of symptotic normality and efficiency is an important reason for the use of M.L.

Change of Variables

Another convenient property of M.L. estimators is as follows. If we change variables from θ to a function of θ, then the M.L. estimate changes to that same function of $\hat{\theta}$.

$$\hat{g}(\theta) = g(\hat{\theta}) \text{ ,because } \frac{\partial \mathscr{L}}{\partial g} = \frac{\partial \mathscr{L}}{\partial \theta}\frac{\partial \theta}{\partial g} ,$$

assuming $\frac{\partial \theta}{\partial g}$ exists, and therefore $\frac{\partial \mathscr{L}}{\partial \theta} = 0$ implies that $\frac{\partial \mathscr{L}}{\partial g} = 0$ at the same θ. There might be additional solutions if $\frac{\partial \theta}{\partial g} = 0$ at any points in the physical region. That will normally not happen if g is a one-to-one function of θ unless points of inflection occur:



Therefore M.L. works like the method of substitution in going to a function of a parameter.

This is another illustration that M.L. estimators are often biased, since if $\hat{\theta}$ is unbiased for θ, $g(\hat{\theta})$ may be biased for $g(\theta)$. However, as $n \rightarrow \infty$, M.L. provides unbiased and efficient estimators for *both* g and θ under the "suitable" conditions listed above. $\mathscr{L}$ approaches the normal shape differently, at a different rate, depending on whether g or θ is the variable being estimated.

For functions g of a multivariate θ, the transformed information matrix is

$$I(g_{k\ell}) = \left(\frac{\partial g_k}{\partial \theta}\right)^T I(\vec{\theta})\frac{\partial g_\ell}{\partial \theta} ;$$

θ and g do not need to have the same dimensions.

More on Errors

As mentioned above, if I actually depends upon θ_{true}, an unknown, we can use $I(\hat{\theta})$ for purposes of estimating the error of an experimental statistic.

It is possible to compute higher-order corrections to improve the estimate of the errors in $\hat{\theta}$ or $\hat{g}$ for finite n.

To estimate the error on the M.L. estimate for finite n, we again use the square root of the variance. Since $\mathscr{L} = \prod_i f(x_i|\vec{\theta})$, it is a joint p.d.f. for $\vec{x}$ (not $\vec{\theta}$), and we may be able to evaluate a covariance matrix as

$$V_{ij}(\theta) = \frac{\int (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)\,\mathscr{L}\,d\vec{x}}{\int \mathscr{L}\,d\vec{x}}$$

(we must normalize since $\mathscr{L}$ is defined only to within a multiplicative constant).

If θ is efficient, and therefore $\mathscr{L}$ takes on the form of a Gaussian, this covariance matrix = $I^{-1}(\hat{\theta})$. Therefore, $(I^{-1}(\hat{\theta}))_{ij}$ gives the value of V_{ij}.

For a single θ:

$$I = E[S^2] = -E\left[\frac{\partial S}{\partial \theta}\right]$$

$$= E\left[(\frac{\partial \ell n \mathscr{L}}{\partial \theta})^2\right] = -E\left[\frac{\partial^2 \ell n \mathscr{L}}{\partial \theta^2}\right].$$

For multidimensional θ:

$$V_{ij}^{-1} = E\left[\frac{\partial \ell n \mathscr{L}}{\partial \theta_i}\frac{\partial \ell n \mathscr{L}}{\partial \theta_j}\right] = -E\left[\frac{\partial^2 \ell n \mathscr{L}}{\partial \theta_i\,\partial \theta_j}\right].$$

The last equality is independent of θ and follows from $\frac{\partial^2}{\partial \theta_i\,\partial \theta_j}\int f dx = \frac{\partial^2}{\partial \theta_i\,\partial \theta_j}(1) = 0$. If the estimator is efficient, then we have exactly

$$-E\left[\frac{\partial^2 \ell n \mathscr{L}}{\partial \theta^2}\right] = -\left.\frac{\partial^2 \ell n \mathscr{L}}{\partial \theta^2}\right|_{\theta=\theta_{true}} = -\left.\frac{\partial^2 \ell n \mathscr{L}}{\partial \theta^2}\right|_{\theta=\hat{\theta}}$$

(the "expectation" operation has no effect, since the score is perfectly linear and therefore the slope of the score is independent of θ).

To estimate this for a real sample, we replace the expectation value by the sample mean and evaluate at $\hat{\theta}$:

$$V_{ij}^{-1} \simeq \frac{1}{n}\sum_i^n \left.\frac{\partial \ell n \mathscr{L}}{\partial \theta_i}\right|_{\hat{\theta}_i} \left.\frac{\partial \ell n \mathscr{L}}{\partial \theta_j}\right|_{\hat{\theta}_j}$$

The expectation form is useful for estimating the error we expect from a measurement, e.g., prior to the experiment, as part of the design study for the experiment. This assumes, of course, that we can make an initial guess at θ when the estimator is not efficient.

We have already derived the variance for our weighted mean M.L. example, using first considerations. From the above, we can estimate the same quantity:

we have $\hat{\mu} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$ and $\frac{\partial \ell}{\partial \mu} = \sum \frac{x_i}{\sigma_i^2} - \sum \frac{\mu}{\sigma_i^2}$;

therefore

$$\frac{\partial^2 \ell}{\partial \mu} = -\Sigma \frac{1}{\sigma_i^2} \, .$$

Therefore

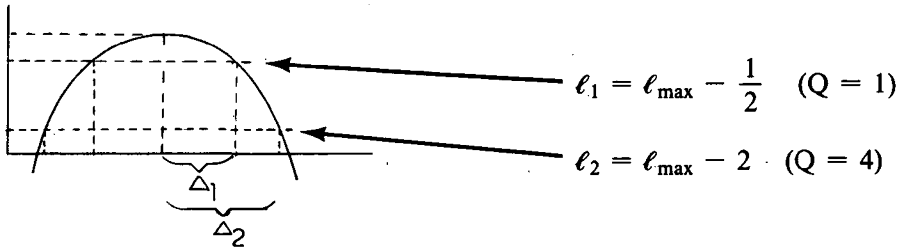
$$-\,\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \mu^2}\right] = \Sigma \frac{1}{\sigma_i^2} = \frac{1}{\mathrm{V}(\hat{\mu})} \, ,$$

which is the same answer we got before. It is independent of the data x_i and therefore $\hat{\mu}$; it depends only upon the data errors σ_i, known in advance. That is, since the estimator $\hat{\mu}$ is efficient, the "expectation" operation had no effect.

Geometrical Method of Estimating Errors

In many problems, the solution to the M.L. problem has to be found numerically. In this event, we can find the errors from the above in the large sample case, where the estimator becomes efficient. In the small sample case, we can make a graphical analysis:

First, go back to the case in which $\mathscr{L}$ is Gaussian. Then $\mathscr{L} = \mathscr{L}(\max)e^{-\frac{1}{2}Q}$, where $Q = (\hat{\theta} - \theta)^2/\sigma^2$; $\sigma^2 = [I(\theta)]^{-1}$; $\ell = \ell n\mathscr{L} = \ell n\mathscr{L}(\max) - \frac{1}{2}Q$. Then the point at which $(\hat{\theta} - \theta)^2 = \sigma^2$, i.e., one standard deviation, can be read off just as if $\mathscr{L}$ were truly a Gaussian p.d.f. for θ:


$\ell_1 = \ell_{\max} - \frac{1}{2} \quad (Q = 1)$
$\ell_2 = \ell_{\max} - 2 \quad (Q = 4)$
Δ_1
Δ_2

Δ_1 = 1 st.d. (standard deviation)

Δ_2 = 2 st.d.

In this case, $\Delta_2 = 2\Delta_1$.

In general, for $\ell = \ell_{\max} - a$ or $\mathscr{L} = \mathscr{L}_{\max}e^{-a}$ (so $a = \frac{1}{2}Q$), the 1, 2, and 3 st.d. confidence intervals may be obtained geometrically from:

$\sqrt{Q}$	a	$P(\lvert\theta - \hat{\theta}\rvert < \Delta_a)$	comment
1	$\frac{1}{2}$	68.3%	$\Delta^2_{\frac{1}{2}} = V(\hat{\theta})$ if $\hat{\theta}$ unbiased
2	2	95.4%	
3	$4\frac{1}{2}$	99.7%	

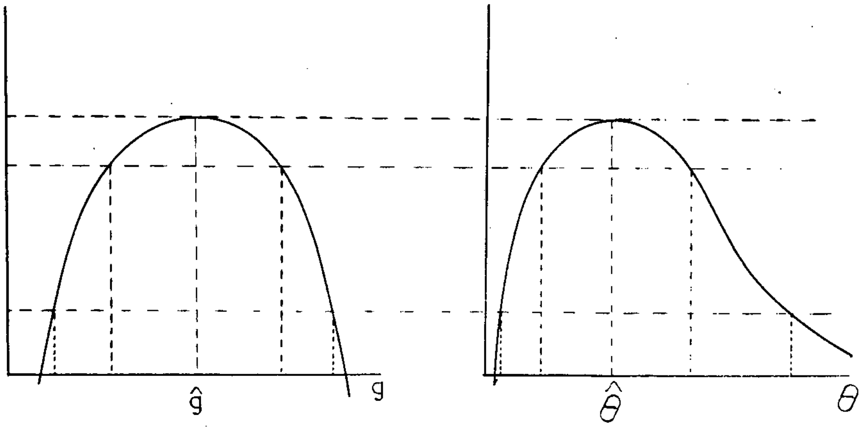
Now, if $\mathscr{L}(\theta)$ is not normal, imagine a one-to-one transformation $g(\theta)$ to a new parameter g such that $\mathscr{L}(g)$ is normal (such a g may not exist — see below). Now $\hat{g}(\theta) = g(\hat{\theta})$ as we have already shown. Let h be g^{-1}, i.e., $\theta = h[g(\theta)]$. Then $P(\lvert g - \hat{g}\rvert < \Delta_g) = P(\lvert\theta - \hat{\theta}\rvert < \Delta_\theta)$ if $g = g(\theta)$, and Δ_θ is the transformed interval, both corresponding to the same a:

$$\Delta_g = g \text{ interval corresponding to } \pm[\ell_{\max}(\hat{g}) - a];\ \Delta_\theta = \theta \text{ interval corresponding to } \pm[\ell_{\max}(\hat{\theta}) - a].$$

Whether as a function of g or of θ, we can estimate the errors if we drop down an amount "a" from the top. We use the same a, since this is not a change of variables — x_i is still the variable — but just a change in parameters, i.e.,

$$\mathscr{L}(\theta \vert x) = \mathscr{L}[g(\theta)\vert x]\,, \text{ for all } \theta.$$

This is a non-trivial point. For fixed $\vec{x}$, $f(\vec{x})$ must be unchanged whether we write the p.d.f. as a function of θ or of g. θ only chooses the member of the p.d.f. family. $g(\theta)$ must choose the same member. The likelihood certainly has a different shape graphed versus g than versus θ, but must have the same value at $g(\theta)$ as at the corresponding parameter θ. The likelihood, a function of the parameters for fixed data x_i, is a function of a *point;* it is not a p.d.f., which is a function of an *interval.* That is why we multiply by a Jacobean when we change variables for a p.d.f. No Jacobean is needed for a likelihood function.



Since $\mathscr{L}$ has the same value at corresponding values of g and θ, we can construct the "errors" in the same way for either one. These may not be symmetric, as shown above. Notice that we can find the errors on $\hat{\theta}$ without actually finding g.

A problem is that no such g may exist. However, it does exist asymptotically. Therefore there probably exists some transformation to a parameter nearly normally distributed. Since we never need to actually perform the transformation, we adopt the hypothesis:

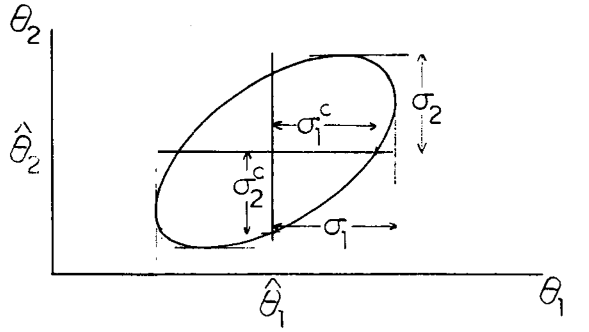
> A Y% confidence interval for θ has limits at values of θ where $\ell_{\max} - \ell = a$, where a and Y correspond as in the table above. This statement means: in a large number of similar experiments, the frequency that a Y% confidence interval constructed according to the above prescription contains the true value of θ is $\cong$ Y%.

In some cases, this hypothesis leads to disjoint intervals, infinite intervals, or other difficulties. Even in ordinary cases, the statement may be only roughly true. If the estimator is not efficient, the information may depend upon θ; since we must estimate it at $\hat{\theta}$, we may make an error in estimating the errors. In any case, this method (or any other method) of estimating errors from the likelihood function should be regarded as a heuristic principle — it is not well grounded in grand mathematical theorems.

Multivariate Case

If all variables are efficient, then the likelihood will be a multivariate normal.

Example: 2 Dimensions



$$\ell n\mathscr{L} = \ell n\mathscr{L}(\max) - \frac{1}{2}Q$$

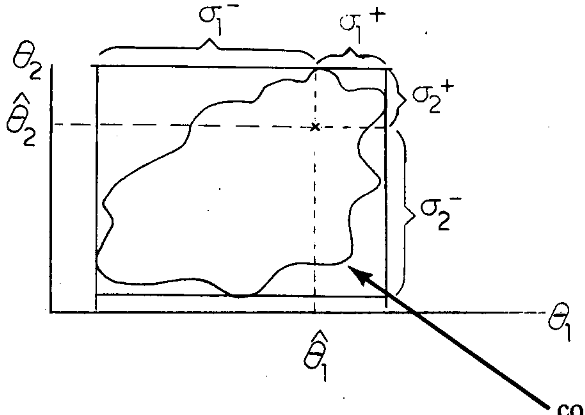
To get the correct error on θ_1, we must allow θ_2 to vary, and vice-versa. Otherwise we get only the reduced (conditional) error σ_1^c or σ_2^c.

Q is the factor in the exponential for the bivariate normal p.d.f. In fact, Q is $\chi^2(2)$, as we have discussed before.

a	$P(\frac{1}{2}Q \leqslant a)$	
0.5	39.3%	$Q = 1^2$
2.0	86.5%	$Q = 2^2$
4.5	98.9%	$Q = 3^2$

as we discussed in the bivariate normal case. In general, for n parameters θ, Q is $\chi^2(n)$, and we can look the corresponding probabilities up in a table of χ^2 c.d.f.'s.

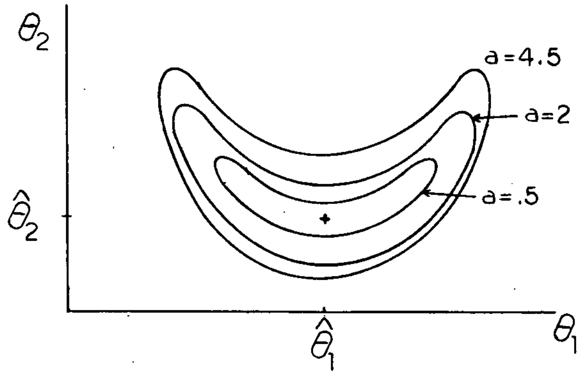
If things are not this regular (the parameter estimators are not all efficient), it is common (there are several approaches possible) to adopt a hypothesis similar to that of the univariate case:



constant $\ell = \ell_{max} - a$; $a = \frac{1}{2}$.

We must solve for these extrema numerically. This must be done for one variable at a time while allowing the other variables to move. Thus, we first guess at a value for θ_1, i.e., $\hat{\theta}_1 + \sigma_1^{guess} = \theta_1^1$, so that θ_1^1 is a guess at the maximum value of θ_1 giving $\ell = \ell_{max} - a$; re-maximize ℓ with θ_2 and check if it is still true that $\ell = \ell_{max} - a$; repeat using techniques of numerical analysis until σ_1^+ is found. Then repeat for σ_1^-, σ_2^+, σ_2^-. We must re-maximize with respect to the other variables at each stage, otherwise we will just find the conditional errors $V(\theta_1|\hat{\theta}_2)$, etc. Then we can define separate confidence intervals for θ_1 and θ_2.

If this box doesn't describe the joint probability reasonably well, then we may want to give the full shape of the likelihood function, e.g.,



Sometimes there is more than one maximum with disjoint regions around each, defined by some set of contours.

Some comments on M.L. estimation:

1) If the sample is large, M.L. gives a unique, unbiased, minimum variance estimate, under certain general conditions. However, "large" is not well defined, and for finite samples, the M.L. estimate may be neither unique, unbiased, nor minimum variance. In this case, other estimators may be preferable. Alternatively, one may begin with the M.L. estimator and make various bias-reducing or variance-reducing corrections.

2) M.L. estimates are often the easiest to compute, especially for complex problems. In many practical cases, M.L. is the only tractable approach.

3) M.L. estimates are sufficient (Chapter 6), that is, they use all the information about θ which is contained in the data. In particular, for small samples they can be very much superior to methods which rely upon binning data, since the binning loses information. Estimators based upon the populations of bins are for this reason not sufficient. Likelihood estimators use every bit of data at face value.

4) M.L. estimates are not necessarily robust. They can sometimes perform very poorly in the presence of deviations from the assumptions. For example, if the true distribution for x_i is a "heavy-tailed Gaussian" and we have assumed a pure Gaussian form, M.L. estimators may have a much larger true variance than some other estimator.

5) M.L. gives no way of testing the validity of the underlying theory, i.e., whether or not the p.d.f. we have assumed is the correct one.

6) The justification for likelihood procedures lies only in the nature of the results, i.e., it usually works. Note that in constructing confidence intervals (errors), we treated the likelihood as if it were a p.d.f. for θ. But this can't be justified theoretically, because, by Bayes' Theorem, if $f(x|\theta)$ is a p.d.f. for x given θ, and $g(\theta|x)$ is a p.d.f. for θ at a given x, then

$$g(\theta|x) \propto f(x|\theta)\, P(\theta) \propto \mathscr{L}(\theta)\, P(\theta) \text{ (for one observation).}$$

To use $\mathscr{L}(\theta)$ as a p.d.f. $g(\theta|x)$ requires assuming $P(\theta) = 1$. See the section on Bayes' Postulate (Chapter 2) for a discussion. Briefly, there is no mathematical justification for this assumption. If $\mathscr{L}(\theta)$ were a p.d.f., we would have been tempted to use its expectation value as an estimator for θ, rather than its maximum.

7) To use M.L., we must know the p.d.f. of our measurements about their true values. This will be seen to be not necessary for the Method of Least Squares; we will only need certain assumptions regarding the p.d.f.

Chapter 11. Constructing Estimators III — Method of Least Squares

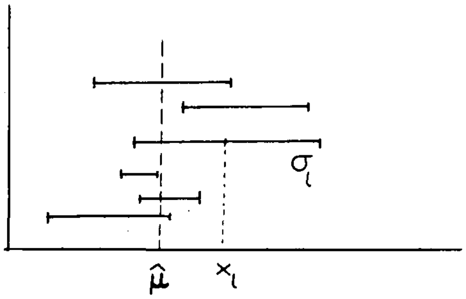
Starting From Maximum Likelihood

In the example of the previous chapter, the case of the sample of x_i, $N(x_i;\, \mu, \sigma_i)$, we constructed

$$\mathscr{L} = \prod_i \frac{1}{\sqrt{2\pi}\,\sigma_i}\, e^{-\frac{(x_i - \mu)^2}{2\sigma_i^2}}$$

$$\ell n \mathscr{L} = C - \sum_i [-\ell n\, \sigma_i - \frac{(x_i - \mu)^2}{2\sigma_i^2}], \quad \text{C a constant.}$$

That is, each x_i is one measurement with known error σ_i of a fixed but unknown quantity μ, which we wish to estimate. An important assumption is that σ_i does not depend upon μ. Otherwise, we must estimate its true value; normally one evaluates it at $\hat{\mu}$, which introduces non-linearities, a subject of later discussion.



The quantity $\frac{|x_i - \mu|}{\sigma_i}$ is one useful way to measure the distance of x_i from μ, namely, in terms of a number of standard deviations. Then any points which are, e.g., k standard deviations away, are weighted equally.

To maximize $\mathscr{L}$, we must minimize the square of this, $\sum_i \frac{(x_i - \mu)^2}{\sigma_i^2}$. This quantity would be $\chi^2(n)$, assuming each point is independent, if μ were known. If μ is unknown, we set $\mu = \hat{\mu}$ and estimate it from the data. Then there is one relationship among the terms of the χ^2, and therefore the quantity

$$Z = \sum \frac{[x_i - \hat{\mu}(x)]^2}{\sigma_i^2} \text{ is } \chi^2\,(n - 1)\,.$$

The method of Least Squares replaces the maximum of the likelihood function with the minimum of this quantity, the sum of the squared distances of the measurements from the final estimate. The minimum value of this χ^2 occurs at

$$\hat{\mu} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}\,, \qquad \text{as we have seen earlier}\,.$$

In this case the method of Least Squares and the Principle of Maximum Likelihood find the same estimator. This would not be true, in general, if the distribution of the data about μ is something other than Gaussian. We will show that, although we have arrived at the Method of Least Squares (L.S.) from the Principle of M.L., L.S. is much more solidly based than M.L. It is also, perhaps as a consequence, less widely applicable.

The result of a M.L. or L.S. procedure is often called a "fit" to the data.

It would also be possible to minimize $\sum|x_i - \mu|$, or some other measure of distance, to find an estimator for μ. We will show that under certain general circumstances the least squares estimator has smallest variance. But if, for example, some of the data have problems such that the assumed values of σ_i are too small (e.g., many data points are in the notorious long tails of the Gaussian), we may desire a more robust estimator. We will not discuss this here, but it can be shown that $\sum|x_i - \mu|$ is minimized by setting $\hat{\mu}$ equal to the median. The median throws away some of the information in the distribution but it is very insensitive to long tails on Gaussians.

We can now substitute $\hat{\mu}$ into the above for Z and calculate the χ^2 we get. If we repeat the experiment a large number of times, the distribution of Z is expected to look like a χ^2 (n $-$ 1), provided the theory is correct, i.e., that the x_i's are normal with true mean equal to μ, and each has true variance σ_i^2. We can make a test of this theory by comparing the value of the χ^2 we get with the χ^2 c.d.f. for n $-$ 1 degrees of freedom. If the value of Z is low enough, this doesn't prove the model, but doesn't rule it out. If the value of Z is larger than some pre-determined, or mutually agreed-upon, value, we say that the data are inconsistent with the model. This value could be "t" such that P(Z > t) is < 1%, or $< \frac{1}{10}$%, or perhaps we don't want to reject the theory unless still lower values are reached. If the theory being tested says "the neutrino has no mass", we want to be very sure before we reject. If the theory says "this track is well described by this curve fitted to these measurements", we can afford to reject at 1% or maybe even 5%, because we know tracks do have kinks sometimes, or that there are sometimes bad measurements.

If we reject the fit, it just means we don't believe the theory describes the present data. This could be just due to some bad data, e.g., outliers at great distance due to some background process — or it could be due to a fundamental error in the theory.

Thus, unlike the general M.L. problem, in this example we have a way of testing for goodness of fit.

If we accept that the data are consistent with the model, we now want to estimate the error in $\hat{\mu}$.

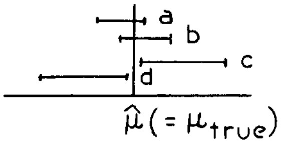
Clearly, we want the error to decrease with increasing n, the number of measurements. We also want the error to increase as the σ_i are increasing, in some way. Do we want the error to increase if the points are farther away, i.e., as the χ^2 increases?

We can estimate the error using propagation of errors (Chapter 9). We are in this example free of the flaws of that technique, because we are starting from a normal variable, x_i, and making a transformation linear in x_i to $\hat{\mu}$:

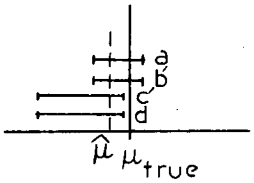
$$V(\hat{\mu}) = \left(\frac{1}{\sum\frac{1}{\sigma_i^2}}\right)^2 \sum\frac{V(x_i)}{\sigma_i^4} = \frac{1}{\sum\frac{1}{\sigma_i^2}} ,$$

as we saw before. This is *independent* of the value of χ^2, which depends upon the particular measurements x_i. That is, the same set of σ_i's, known before the fit is done, will result in the same $V(\hat{\mu})$

regardless of where the data points lie or what the final $\hat{\mu}$ is or the final χ^2 is. $V(\hat{\mu})$ is also independent of μ for the same reason. We will see later that if we had to estimate σ^2 as well as $\hat{\mu}$, *then* the errors would depend on the value of the χ^2. In the meantime, we can qualitatively understand the fact that the error should not depend upon the χ^2 as follows. Assume the following four measurements, which happen to find $\hat{\mu}$ = the true value of μ:



The above, which predicts the correct μ_{true}, has the same errors as this case:



(Through an accidental fluctuation, the points cluster better about some value than they did above, but we don't know that it's the true value.)

which predicts $\hat{\mu}$. The second case has the same real probability of occurrence as the first case (b′ and c′ are reflected about μ_{true} and have exactly the same true error as b and c), but has lower χ^2 (about $\hat{\mu}$, remember); we don't know the true value of μ any better in the second case than in the first case.

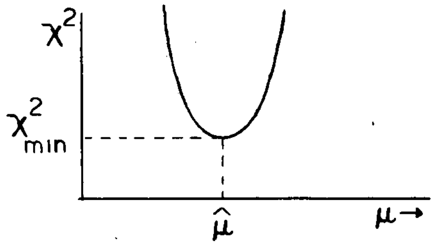
But we note that the variance of $\hat{\mu}$ depends upon the *shape* of the χ^2:

$$Z = \sum \frac{(x_i - \hat{\mu})^2}{\sigma_i^2}$$

$$\frac{\partial Z}{\partial \hat{\mu}} = -\sum \frac{2(x_i - \hat{\mu})}{\sigma_i^2} = 0 \text{ by choice of } \hat{\mu}\,.$$

$$\frac{\partial^2 Z}{\partial \hat{\mu}^2} = +2\sum \frac{1}{\sigma_i^2} = 2/V(\hat{\mu})\,.$$

All higher derivatives = 0, a consequence of the efficiency of the estimator. Thus the χ^2 is a perfect parabola:



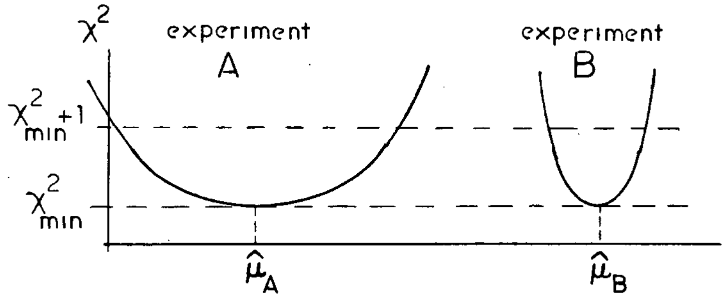
$$\chi^2(\mu) = Z(\mu) = Z(\hat{\mu}) + \frac{2}{2}\,\frac{(\hat{\mu} - \mu)^2}{V(\hat{\mu})}$$

Corresponding to what we did for the likelihood case, for this way of looking at the same problem we can construct the errors in $\hat{\mu}$ by drawing lines of constant χ^2. For a given value of μ,

$$\sigma^2_{\hat{\mu}} = V(\hat{\mu}) = \frac{(\hat{\mu} - \mu)^2}{Z(\mu) - Z(\hat{\mu})} .$$

In particular, if we can find values of μ such that $Z(\mu) = Z(\hat{\mu}) + 1$, then $(\hat{\mu} - \mu)^2 = \sigma^2_{\hat{\mu}}$ and $(\hat{\mu} - \mu) = \pm\sigma_{\hat{\mu}}$.

Therefore ±1 standard deviation is observed at the points where the horizontal line $Z(\mu) = Z(\hat{\mu}) + 1$ intersects the χ^2 parabola. $\pm 2\sigma$ occurs at $Z(\mu) - Z(\hat{\mu}) = 4$, etc.



The estimator for μ_B is more tightly determined, because the χ^2 gets worse much more rapidly as we go away from $\hat{\mu}_B$.

We will now look at more general cases, and also see what we can learn if the measurements are not Gaussian but still unbiased.

The Two-Variable Case: the Linear Model

In the preceding example, we made a number of measurements of a fixed quantity. Now let us suppose we make a number of measurements of a quantity y which varies with some characteristic x of the measurement. We will assume for now that x (position, time of day, amount of a chemical, price of an item, etc.) is chosen by the experimenter and is known without error. At each x_i we find a measured y_i which has known expected error σ_i. The variable x is called the *independent* variable or *predictor* variable, and y is called the *dependent* variable or *response* variable. We will assume σ_i is a constant which does not depend upon y at any given x.

One of the reasons for doing a fit to a curve is to enable us to predict the most likely value of *future* measurements at a specified x. For example, we wish to calibrate an instrument, such as a thermometer. Then the predictor variable x would be the value the instrument reads. The response variable y would be the actual value (here, the actual temperature). We do a curve fit in order to average out the fluctuations of individual data points as much as possible. This only works, of course, if the model used for the curve is at least approximately correct. A more general example of this type of problem is any case in which the y_i's themselves, rather than the θ's (the parameters of the curve), are the parameters of interest. In this case, we improve our understanding of the y_i by using $\hat{y}_i$ in place of y_i, provided we believe our model is correct.

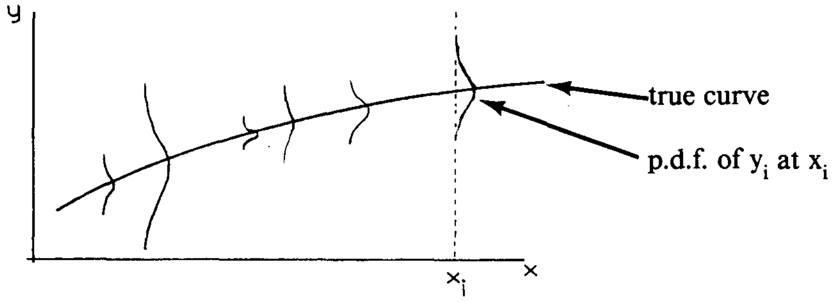
We will assume that we have a model for y versus x in terms of certain parameters θ which are coefficients of known functions of x, i.e., of the following form:

$$y = \theta_1 h_1(x) + \theta_2 h_2(x) + \cdots + \theta_k h_k(x) .$$

This is the curve to which we fit. There are "k" parameters to be estimated. The important features of this model are that the h_i's are *any* known (frequently, one of them is a constant), distinguishable, functions of x, single-valued over the allowed range of x, and that y is *linear* in the θ_i's. Our task is to estimate the θ_i's. Therefore, this is called a linear model, even though y is not necessarily linear in x. In some cases, the linear model is just an approximation arrived at by retaining only the first few terms of a Taylor series. None of the h_i's may be expressible as a linear combination of any set of the others; otherwise, the corresponding θ_i will be indeterminate.

We wish to determine the values θ_i such that the model provides the best "fit" to the measurements. We assume that any deviations of y_i from this curve must be due to random measurement error or some other unbiased effects beyond our control, but whose distribution is known from previous study of the measuring process to have variance σ_i^2. It need not be Gaussian. We will take as our measure of the distance of y_i from the hypothesized curve the squared distance in units of σ_i, as above.

The general term for this fitting procedure is Regression Analysis. This term is of historical origin and is used even though nothing is thought to "regress".



We assume that our actual measurements are described by

$$y_i = \sum_{j=1}^{k} \theta_j h_j(x_i) + \epsilon_i ,$$

where the unknown error ϵ_i has the properties: $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma_i^2$ is known (and sometimes ϵ_i is Gaussian). We will state when we need the Gaussian assumption. We don't need that assumption for most of what we will do; only after the fit when we test goodness of fit. Note: If at each x_i, the y_i are not normal, we may be able to transform to a set of variables which is, for when we need the Gaussian model. We are implicitly assuming that each y_i is an independent measurement; if the errors on the y_i are correlated we will need a non-diagonal $m \times n$ error matrix, as will be discussed. The x_i may be chosen any way we wish, including on top of one another. We shall see, however, that we will need at least k distinct x_i's to determine k parameters θ_i. We will continue to denote by "n" the number of measurements (x_i, y_i).

The deviations from the true curve are ϵ_i; therefore our χ^2 is

$$\chi^2 = \sum_{i=1}^{n} \epsilon_i^2/\sigma_i^2$$

(This is not a true χ^2, i.e., it does not come from a χ^2 p.d.f., unless the ϵ_i are normal [see Chapter 3]. In some references this is called "SS" for "Sum of Squared deviations.")

$$= \sum_i \frac{1}{\sigma_i^2} [y_i - \sum_{j=1}^{k} \theta_j h_j(x_i)]^2 \, .$$

We don't know the actual value of this, since we don't know the actual values of the elements of $\vec{\theta}$. The L.S. method tells us to estimate $\vec{\theta}$ by that $\hat{\theta}$ which minimizes χ^2. Hence, writing

$$\frac{\partial \chi^2}{\partial \theta_\ell} = 2 \sum_i \frac{1}{\sigma_i^2} [y_i - \sum_j \theta_j h_j(x_i)][- h_\ell(x_i)]$$

$$= 0$$

for all $\ell = 1, \ldots , k$ provides the solution for the k components of $\hat{\theta}$ at minimum χ^2. Therefore

$$\sum_i \frac{h_\ell(x_i)}{\sigma_i^2} \sum_j \hat{\theta}_j h_j(x_i) = \sum_i \frac{y_i}{\sigma_i^2} h_\ell(x_i) \, .$$

These are called the *normal equations* for $\hat{\theta}_j$. Their solution is the L.S. solution to the problem. It's a bit easier in matrix notation, because these are a set of k linear equations in k unknowns:

$$\text{Write } \vec{y} = \underset{(n \times 1)}{\begin{pmatrix} y_1 \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ y_n \end{pmatrix}} \; ; \; \vec{\theta} = \underset{(k \times 1)}{\begin{pmatrix} \theta_1 \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \theta_k \end{pmatrix}} \; ; \; \vec{\epsilon} = \underset{(n \times 1)}{\begin{pmatrix} \epsilon_1 \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \epsilon_n \end{pmatrix}} \; . \quad \leftarrow \text{dimensionality}$$

It is customary in L.S. analysis to define vectors as being column vectors. The matrix we need is:

$$\mathbf{H} = \begin{pmatrix} h_1(x_1) & h_2(x_1) \; .... \; h_k(x_1) \\ & \bullet \\ & \bullet \\ & \bullet \\ & \bullet \\ h_1(x_n) & h_2(x_n) \; .... \; h_k(x_n) \end{pmatrix} \qquad (n \times k) \quad .$$

$$\text{Then } \mathbf{H}\, \vec{\theta} = \begin{pmatrix} \sum_j \theta_j \, h_j \, (x_1) \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \sum_j \theta_j \, h_j \, (x_n) \end{pmatrix} \qquad (n \times 1) \quad ,$$

and our model is

$$\vec{y} = \mathbf{H}\,\vec{\theta} + \vec{\epsilon}\ .$$

Since $E(\vec{\epsilon}) = 0$, we obtain $E(\vec{y}) = \mathbf{H}\,\vec{\theta}$, assuming the model is correct. This is just the statement that the expectation value of each measurement lies exactly on the theoretical curve.

The errors σ_i^2 may be incorporated into a variance matrix:

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \ldots & 0 \\ \cdot & & \cdot \\ \dot{0} & \cdots & \dot{\sigma}_n^2 \end{bmatrix}$$

(n × n) symmetric. If we approximate this by $\mathbf{V} = \mathbf{I}\sigma^2$ (i.e., fixed σ^2 for all measurements), we still get an unbiased estimator, but not minimum variance (unless all the σ_i^2 are really equal, of course).

Note that $\mathbf{H}\vec{\theta}$ is fixed (once the data have been taken), therefore $\mathbf{V}(\vec{y}) = \mathbf{V}(\vec{\epsilon})$, which is $\mathbf{V}$. We are talking about the variance in y due to measurement error about the true curve. Note that we can drop the assumption of independence of the measurements here, if we want, by using non-zero off-diagonal V_{ij}'s. Now our χ^2 (or "SS") is

$$\chi^2 = \vec{\epsilon}^{\,T}\,\mathbf{V}^{-1}\,\vec{\epsilon} \quad (1 \times 1)$$

$$= (\vec{y} - \mathbf{H}\vec{\theta})^T\,\mathbf{V}^{-1}\,(\vec{y} - \mathbf{H}\vec{\theta})\ .$$

Therefore

$$\underset{(k\times 1)}{\frac{\partial\chi^2}{\partial\theta}} = -\,2\mathbf{H}^T\mathbf{V}^{-1}\,(y - \mathbf{H}\vec{\theta}) = 0 \text{ implies}$$

$$\underset{(k\times n)}{\mathbf{H}^T}\ \underset{(n\times n)}{\mathbf{V}^{-1}}\ \underset{(n\times 1)}{\vec{y}} = \underset{(k\times n)}{\mathbf{H}^T}\ \overset{(k\times k)}{\underset{(n\times n)}{\mathbf{V}^{-1}}}\ \underset{(n\times k)}{\mathbf{H}}\ \underset{(k\times 1)}{\hat{\theta}}\ , \quad (k\times 1)$$

the normal equations, which may be formally solved by inverting the square matrix $\mathbf{H}^T\,\mathbf{V}^{-1}\,\mathbf{H}$. This matrix is symmetric, since V is symmetric:

$$\hat{\theta} = (\mathbf{H}^T\,\mathbf{V}^{-1}\,\mathbf{H})^{-1}\,\mathbf{H}^T\,\mathbf{V}^{-1}\,\vec{y}\ .$$

Note: $\hat{\theta}$ is a (k×1) vector, even though we don't have room for the "→" above it.

It is useful to note at this point that the actual sizes of the errors don't need to be known to find $\hat{\theta}$; only their *relative* sizes. Thus, if we write $\mathbf{V} = \sigma^2\mathbf{W}$, where σ^2 provides the dimension and scale and $\mathbf{W}$ just provides dimensionless relative weights, we can see that

$$\hat{\theta} = (\mathbf{H}^T\,\mathbf{W}^{-1}\,\mathbf{H})^{-1}\,\mathbf{H}^T\,\mathbf{W}^{-1}\,\vec{y}\,,$$

with no mention of σ^2. We will, of course, need to know the scale σ^2 to estimate the errors on $\hat{\theta}$.

Note:

$$E(\hat{\theta}) = E[(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{V}^{-1}\vec{y}] = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{V}^{-1}E(\vec{y}) = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})\,\vec{\theta} = \vec{\theta}\,.$$

Therefore $\hat{\theta}$ is unbiased for $\vec{\theta}$, always assuming the model is correct.

Procedures exist for solving the normal equations without the intermediate step of taking the inverse.

In some cases, it is more convenient to solve these equations by numerical approximation methods rather than the exact formulation above. Very clever programs exist as pre-packaged routines for finding the minimum of a χ^2 (or any other) function. One usually codes a subroutine which calculates χ^2 for a given set of trial values of the parameters. The program needs a set of starting values supplied by the experimenter. It evaluates the χ^2 at numerous points in $\vec{\theta}$ space, determines the most likely direction in this space to find the minimum, and proceeds to search until the minimum is found. If analytical derivatives are available, more rapid convergence can usually be attained. For simple cases like the linear problem we have considered here, such a numerical approximation method is not very wasteful of computer time, and its simplicity decreases the probability of an experimenter's error and usually saves him or her time as well. If the problem is not linear, as in cases we shall discuss, numerical approximation techniques, with or without explicit calculation of derivatives by the experimenter, are often the method of choice.

If there exists a linear relationship among some of the h_j's, then the columns of **H** are not all independent, and since **V** is symmetric, $\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H}$ will be singular. The best step then is to eliminate some of the h's and their associated θ's until a solution can be found. Also, there must be at least k distinct x's, or the same term will be singular.

Note that if n = k, and there are only as many (distinct) x's as there are parameters to be estimated, then

$$(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1} = \mathbf{H}^{-1}\mathbf{V}(\mathbf{H}^T)^{-1} \qquad (k = n) \ ,$$

and $\hat{\theta} = \mathbf{H}^{-1}\vec{y}$ independent of the errors, since the curve will then try to go through all the points [if the above conditions are satisfied to provide non-singular $(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})$]. However, it may not be *able* to go through all the points; for example, if there is no constant term in the h's.

What is the error of our estimator? We want the covariance matrix:

$$\overset{(k\times k)}{\mathbf{V}(\hat{\theta})} = \overbrace{[(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\,\mathbf{H}^T\mathbf{V}^{-1}]}^{(k\times n)}\ \overset{(n\times n)}{\mathbf{V}(\vec{y})}\ \overbrace{[(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\,\mathbf{H}^T\mathbf{V}^{-1}]^T}^{(n\times k)} \ .$$

This can be demonstrated by working out a simple example. Alternatively, it follows from propagation of errors (Chapter 9), since we are converting from $\vec{y}$ to $\vec{\theta}$, and

$$D_i = D(\hat{\theta}_i) = \begin{pmatrix} \dfrac{\partial\hat{\theta}_i}{\partial y_1} \\ \bullet \\ \bullet \\ \bullet \\ \dfrac{\partial\hat{\theta}_i}{\partial y_n} \end{pmatrix} \quad ; \quad \frac{\partial\hat{\theta}_i}{\partial y_j} = [(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{V}^{-1}]_{ij}$$

and $\mathbf{V}(\hat{\theta}) = \mathbf{D}^T\mathbf{V}(y)\mathbf{D}$, hence the above form.

Proceeding, we note that $\mathbf{V}(\vec{y}) \equiv \mathbf{V}$. Also $\mathbf{V}^{-1}$ is square symmetric; therefore $(\mathbf{V}^{-1})^T = \mathbf{V}^{-1}$.

Similarly $[(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}]^T = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}$. Therefore

$$\mathbf{V}(\hat{\theta}) = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}$$

$$= \underset{I}{(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H}}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1} \text{. Therefore}$$

$$\mathbf{V}(\hat{\theta}) = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1} .$$

Derivative Formulation

We note that

$$\chi^2 = (\vec{y} - \mathbf{H}\hat{\theta})^T\,\mathbf{V}^{-1}\,(\vec{y} - \mathbf{H}\hat{\theta}) \ ,$$

$$\left.\frac{\partial\chi^2}{\partial\hat{\theta}}\right|_{\theta=\hat{\theta}} = -\,2\,\mathbf{H}^T\,\mathbf{V}^{-1}\,(\vec{y} - \mathbf{H}\hat{\theta}) \ \text{, and}$$

$$\left.\frac{\partial^2\chi^2}{\partial\theta^2}\right|_{\theta=\hat{\theta}} = +\,2\,\mathbf{H}^T\,\mathbf{V}^{-1}\mathbf{H} = 2\,\mathbf{V}^{-1}\,(\hat{\theta}) \quad \text{(a constant)} \ ,$$

a very useful way to calculate this same matrix. We established the same result for the earlier simple case of repeated measurements of a fixed quantity. In these relations, note that $\frac{\partial\chi^2}{\partial\theta}$ is a (k×1) vector whose i^th element is $\frac{\partial\chi^2}{\partial\theta_i}$ and $\frac{\partial^2\chi^2}{\partial\theta^2}$ is a (k×k) matrix whose ij^th element is $\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}$.

We can, in fact, re-write the solution $\hat{\theta}$ completely in terms of derivatives of the χ^2 without the necessity of constructing **H**, **V**, or the associated matrix products. Begin with

$$\hat{\theta} = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\,\mathbf{H}^T\mathbf{V}^{-1}\hat{y} \ .$$

Now evaluate the χ^2 and its derivatives as above, except at some convenient point $\vec{\theta}_0$, since we don't yet have $\hat{\theta}$. This point is usually chosen as close to the location of the final solution as possible (i.e., we try to make a reasonable first guess), in case the linearity of the problem is only an approximation valid over a small range. Then

$$\hat{\theta} = 2\left(\frac{\partial^2\chi^2}{\partial\theta^2}\right)^{-1}\mathbf{H}^T\mathbf{V}^{-1}\vec{y}$$

$$= \left(\frac{\partial^2\chi^2}{\partial\theta^2}\right)^{-1}\left[\left(\frac{\partial^2\chi^2}{\partial\theta^2}\right)\cdot\vec{\theta}_0 - \frac{\partial\chi^2}{\partial\theta}\right] \quad \text{(all derivatives evaluated at } \vec{\theta}_o\text{)}$$

$$= \vec{\theta}_0 - \left(\frac{\partial^2\chi^2}{\partial\theta}\right)^{-1}\cdot\frac{\partial\chi^2}{\partial\theta} \ .$$

This formula expresses the Newton-Raphson method for finding the solution of the equation $\frac{\partial\chi^2}{\partial\theta} = 0$. The solution is exact in the linear problem because the form of the χ^2 is exactly a parabola, i.e., $\frac{\partial^2\chi^2}{\partial\theta^2}$ is a constant. As we shall see, in a non-linear problem in which $\vec{\theta}_0$, the starting value, is sufficiently close to the final solution, the solution may be found by repeating the procedure substituting $\hat{\theta}$ as the value of $\vec{\theta}_o$ for the next iteration. One terminates the iterations when some suitable convergence criterion is satisfied, perhaps involving the size of the change in $\hat{\theta}$ from the last iteration, or the size of the derivative, which is 0.0 at solution.

The derivative formulation for the L.S. solution is frequently the most convenient technique in practical problems. We note that, by the linearity assumption,

$$\frac{\partial \chi^2}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \sum_k \epsilon_k^2/\sigma_k^2 = 2\sum_k \frac{\epsilon_k}{\sigma_k^2} \frac{\partial \epsilon_k}{\partial \theta_i}$$

and

$$\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} = 2\sum_k \frac{1}{\sigma_k^2} \left(\frac{\partial \epsilon_k}{\partial \theta_i}\right) \left(\frac{\partial \epsilon_k}{\partial \theta_j}\right),$$

since $\dfrac{\partial^2 \epsilon_k}{\partial \theta_i \partial \theta_j}$ vanishes. Also, $\dfrac{\partial \epsilon_k}{\partial \theta_i} = -h_i(x_k)$. Thus the derivatives are easy to compute.

Finally, the value of the χ^2 at the solution, useful for testing our belief in the model as we shall see, is

$$\chi^2(\hat{\theta}) = \chi^2(\vec{\theta}_0) + \left.\frac{\partial \chi^2}{\partial \theta}\right|_{\theta_0} \bullet (\hat{\theta} - \vec{\theta}_0) + \frac{1}{2}(\hat{\theta} - \vec{\theta}_0)^T \left.\frac{\partial^2 \chi^2}{\partial \theta^2}\right|_{\theta_0} (\hat{\theta} - \vec{\theta}_0) .$$

One must be careful to observe the dot and matrix products, as indicated.

Just as in the first example of this chapter, we can show, by expanding χ^2 in this same way but about the solution $\hat{\theta}$, that the set of values $\vec{\theta}$ such that

$$\chi^2(\vec{\theta}) = \chi^2_{\min} + 1 \qquad [\chi^2_{\min} \equiv \chi^2(\hat{\theta})]$$

defines one standard deviation errors in the same way that the geometrical method found the same errors in maximum likelihood analysis (Chapter 10). More on this later.

So far, we have not used the Gaussian nature of ϵ_i. All the results so far therefore apply for any ϵ_i satisfying the simple conditions $E(\epsilon_i) = 0$; $V(\epsilon_i) = \sigma_i^2$ known.

Gauss-Markov Theorem

We state without proof the Gauss-Markov Theorem, which provides L.S. analysis with the firm mathematical foundation lacking in M.L. analysis. If $E(\epsilon_i) = 0$ and $V(\epsilon_i)$ is finite and fixed (V doesn't have to be diagonal), then the least squares estimate $\hat{\theta}$ is unbiased and has uniformly minimum variance among all *linear* unbiased estimates, regardless of the p.d.f. for the ϵ_i. To estimate the errors correctly $V(\epsilon)$ must be known.

Notes:

a) It may be possible to do better than this by using non-linear unbiased estimators.

b) Least Squares does not in general give the same result as M.L. (unless the ϵ_i are Gaussian) even for linear models. In this case, linear least squares fits are often to be preferred to linear M.L.

fits where applicable and convenient. An exception may occur in small sample cases where the data must be binned in order to do least squares analysis, causing a loss of information, and in certain other cases.

Let us work a simple example which will lead us into the study of errors and residuals. We will use this example throughout, but state more general results also, as we go along.

A Straight-Line Fit

Model: $y = a + bx; \ \theta = \binom{a}{b}; \ \mathbf{H} = \begin{pmatrix} 1 & x_1 \\ & \bullet \\ & \bullet \\ & \bullet \\ 1 & x_n \end{pmatrix}$ $(n \times 2)$

$$y_i = a + bx_i + \epsilon_i$$

$$\mathbf{V}(y_i) = \mathbf{V}(\epsilon_i) = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \cdot & & \cdot \\ \dot{0} & \cdots & \dot{\sigma}_n^2 \end{pmatrix}$$ (i.e., uncorrelated measurements)

$$\chi^2 = \sum_{i=1}^{n} \epsilon_i{}^T V \epsilon_i$$

Note: This is not a true χ^2 unless the ϵ_i are Gaussian.

$$= \sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_i^2} = \sum_i (y_i - a - bx_i)^2/\sigma_i^2$$

$$\frac{\partial\chi^2}{\partial a} = 0 \text{ and therefore } \hat{a} = \frac{1}{\Sigma \frac{1}{\sigma_i^2}} \Sigma \left(\frac{y_i - \hat{b}x_i}{\sigma_i^2} \right)$$

$$\frac{\partial\chi^2}{\partial b} = 0 \text{ and therefore } \hat{b} = \frac{1}{\Sigma \frac{x_i^2}{\sigma_i^2}} \Sigma \frac{x_iy_i - \hat{a}x_i}{\sigma_i^2}$$

Note: If $\sigma_i^2 = \sigma^2$ for all i, neither of these terms depends on σ^2. In general, as we have seen, we could write $\mathbf{V} = \sigma^2\mathbf{W}$ and eliminate the scale σ^2 entirely.

Solving, we find

$$\hat{b} = \frac{\left(\Sigma \frac{x_iy_i}{\sigma_i^2}\right)\left(\Sigma \frac{1}{\sigma_i^2}\right) - \left(\Sigma \frac{y_i}{\sigma_i^2}\right)\left(\Sigma \frac{x_i}{\sigma_i^2}\right)}{\left(\Sigma \frac{x_i^2}{\sigma_i^2}\right)\left(\Sigma \frac{1}{\sigma_i^2}\right) - \left(\Sigma \frac{x_i^2}{\sigma_i^2}\right)}$$

Remember: σ_i is an error in y, not in x, which has negligible error.

We could now substitute back for â, but let's simplify for didactic purposes:

let $\sigma_i = \sigma$ for all i. Therefore $\mathbf{V}(y) = \sigma^2\mathbf{I}$.

$$\text{Therefore } \hat{b} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \quad \text{(exercise for the reader)}$$

$$\hat{a} = \sum_i (y_i - \hat{b}x_i) = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i y_i)(\sum x_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

$$= \overline{y} - \hat{b}\overline{x} \quad . \text{ The overbar refers to the mean.}$$

Notice that we have to have at least two of the x_i's different or the denominators collapse to 0. This is one of the conditions we discussed earlier for a non-singular $(\mathbf{H}^T\,\mathbf{V}^{-1}\,\mathbf{H})$ [i.e., there must exist at least two distinct x_i, since we are estimating two parameters; $\vec{\theta} = (a,b)$].

A more general comment: when the h's are polynomials, the round-off error in the computations often becomes serious if the degree of the polynomials becomes larger than, say, 6 or 7. One approach to help is to re-write the problem in terms of orthogonal polynomials so that the matrix $(\mathbf{H}^T\,\mathbf{V}^{-1}\,\mathbf{H})$ becomes diagonal (if **V** is diagonal), and therefore the matrix inversion is trivial. Most of the round-off error comes from the various matrix operations, including, principally, the inversion. Similar considerations apply elsewhere — whatever functions we are using, things are numerically more tractable if we can diagonalize $\mathbf{H}^T\,\mathbf{V}^{-1}\,\mathbf{H}$ by choosing orthogonal functions. The same is true for the non-linear case.

Errors

$$V(\hat{\theta}) = (\mathbf{H}^T\,\mathbf{V}^{-1}\,\mathbf{H})^{-1} = (\mathbf{H}^T\,\mathbf{H})^{-1}\,\sigma^2$$

$$= \sigma^2 \left[\begin{pmatrix} 1 & \ldots & 1 \\ x_i & \ldots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet \\ 1 & x_n \end{pmatrix} \right]^{-1} = \sigma^2 \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1}$$

$$= \frac{\sigma^2}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} = \frac{\sigma^2}{n\sum(x_i - \overline{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

Therefore

$$V(\hat{a}) = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \overline{x})^2} \; ; \; V(\hat{b}) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2} \; ; \; \mathrm{Cov}(\hat{a},\hat{b}) = -\frac{\sigma^2\overline{x}}{\sum(x_i - \overline{x})^2} \; .$$

It is possible to design our experiment in advance such that the points are chosen to minimize these error terms. Here, we clearly want to maximize the denominator, which is done by taking data only at the two extrema of the possible range of x. This is fine, but it reduces our ability to test "goodness of fit" to nil, as we shall see.

If the y_i's are Gaussian, then $\hat{\theta}$ is also, since a linear transformation of a multivariate normal is also a multivariate normal.

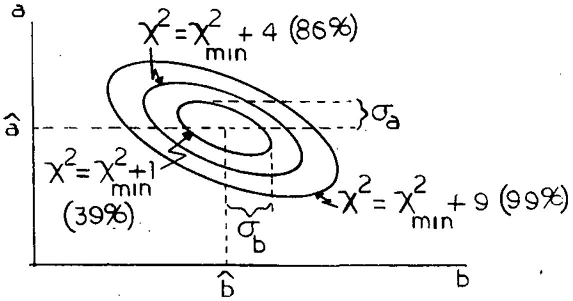
If we plug $\hat{a}$ and $\hat{b}$ back into the equation, we get $\hat{y}_o = \hat{a} + \hat{b}x_o = \overline{y} + \hat{b}(x_o - \overline{x})$ as the predicted value of y at some point x_0 [note that this line passes through $(\overline{x}, \overline{y})$] and $\mathbf{V}(\hat{y}_o) = \mathbf{V}(\hat{a}) + x_o^2\mathbf{V}(\hat{b}) + 2x_o\mathrm{Cov}(\hat{a},\hat{b}) = \mathbf{V}(\overline{y}) + (x_o - \overline{x})^2\mathbf{V}(\hat{b}) + 2(x_o - \overline{x})\,\mathrm{Cov}(\overline{y},\hat{b})$. If $\mathbf{V}(y) = \sigma^2\mathbf{I}$ then $\mathrm{Cov}(\overline{y},\hat{b}) = 0$. [In the general linear case, $\hat{y} = \mathbf{H}_o\hat{\theta}$ is the predicted response and $\mathbf{V}(\hat{y}) = \mathbf{H}_o\mathbf{V}(\hat{\theta})\mathbf{H}_o^T = \mathbf{H}_o(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}_o^T$ ($\mathbf{V} = \mathbf{V}(\overline{y})$, variance of the individual parts) by reversing the same propagation of errors analysis as before]. $\mathbf{V}(\hat{y}_o)$ is the variation expected for the predicted response, which may be much less than that of any measurement we could have made. Thus, if we believe our linear model, then we get a smaller error for y_0 at a given value of x, say x_0, if we measure a lot of responses at x_i's varying around — on both sides of — x_0 and use $\hat{y}_o$ as predicted by the line, rather than if we take one measurement of y_0 right at x_0 and use that. This does depend upon the linear model being correct. Just how much the error is reduced depends upon $\sum(x_i - \overline{x})^2$ in the denominator. The more spread out the measurements are, the more leverage we get and the more well-determined $\hat{a}$ and $\hat{b}$ will be.

Rewriting, assuming $\mathbf{V}(y) = \sigma^2\mathbf{I}$,

$$\mathrm{V}(\hat{y}_o) = \frac{\sigma^2}{n} + \frac{(x_o - \overline{x})^2\sigma^2}{\sum(x_i - \overline{x})^2} \quad \text{(exercise for the reader).}$$

Therefore, the closer x_0 is to $\overline{x}$, the smaller the error in $\hat{y}_o$ will be. If we are right at $\overline{x}$, there is no contribution to $\mathrm{V}(\hat{y}_o)$ from the uncertainty in the slope; all the uncertainty comes just from the uncertainty in $\hat{a}$, which equals the uncertainty in the average y at that point.

If the y_i are normally distributed, then so are the $\hat{\theta}$, in the general case, and the quantity referred to earlier is a true χ^2. We can draw contours of constant χ^2 for a and b



These are ellipses, negatively correlated. They are contours of a bivariate normal (Chapter 3).

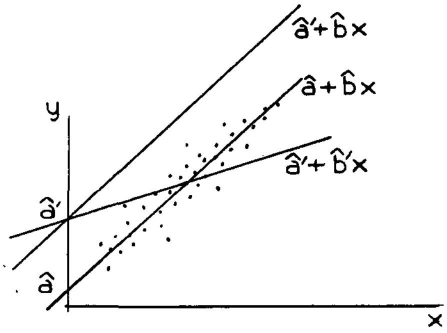
$$\sigma_{\hat{a}}^2 = \mathrm{V}(\hat{a}) = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}_{11}\ ;\ \sigma_{\hat{b}}^2 = \mathrm{V}(\hat{b}) = (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}_{22}\ .$$

It is sometimes of interest to know the distance from the best estimate to the "one-standard-deviation" ellipse, $\chi^2 = \chi^2_{min} + 1$, parallel to one of the axes:

$$(\delta_a)^2 = \mathrm{V}(a|b = \hat{b}) = \frac{\sigma^2}{n} = \frac{1}{(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})_{11}} \leqslant (\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}_{11}$$

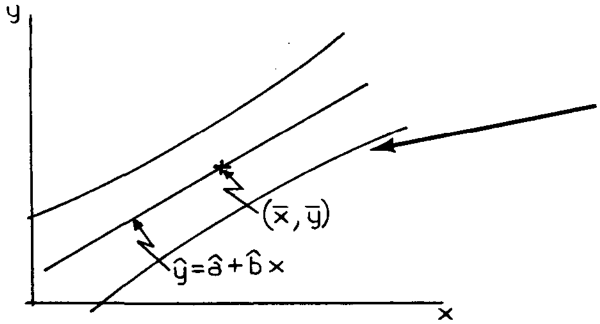
$$(\delta_b)^2 = V(b|a = \hat{a}) = \frac{\sigma^2}{n\overline{x}} = \frac{1}{(H^TV^{-1}H)_{22}} \leqslant (H^TV^{-1}H)_{22}^{-1} \ .$$

Note: The notation $(H^TV^{-1}H)_{ii}^{-1}$ refers to the ii^th^ element of $(H^TV^{-1}H)^{-1}$. Note that the correlation between â and b̂ is negative as in the above figure only if $\overline{x} > 0$, otherwise it is positive. Physically we can draw the following picture to understand this correlation:



If we want to increase â to â′, a straight line with the same slope b̂ quickly becomes a bad fit. But we can achieve a much better fit by decreasing b̂ to b̂′ as shown, so the line still passes by the bulk of the points.

If we convert these errors into errors in ŷ, we get a picture like this


y
$(\overline{x}, \overline{y})$
$\hat{y} = \hat{a} + \hat{b}x$
x

hyperbola: $y = \hat{y} \pm \left\{ \frac{\sigma^2}{n} + \frac{(x - \overline{x})^2\sigma^2}{\sum(x_i - \overline{x})^2} \right\}^{1/2}$,

which corresponds to sliding â up and down by δ_a or twisting the slope b̂ thru $\pm\delta_b$, or a combination of these motions on some other part of the $\chi^2_{min} + 1$ ellipse.

The first term in the equation for the hyperbola corresponds to varying â only, leaving b̂ fixed. Since this just changes the "y" of the line at each x_i by an equal amount, the error in this direction is $\sigma/\sqrt{n}$, the error in $\overline{y}$ of the data points. The second term corresponds to pivoting the line about the point $(\overline{x},\overline{y})$, which requires a simultaneous change in â and b̂. At a given x, one standard deviation due to such pivoting corresponds to $\pm \frac{(x - \overline{x})\sigma}{\sqrt{\sum(x_i - \overline{x})^2}}$. The errors of these two modes of motion add in quadrature.

This error envelope, represented by the bounding hyperbolas, is larger at the ends than at the middle because the curve is more uncertain there, due to the error in the slope. The true line is 39% likely to fall within these bounds. The hyperbola corresponding to a different error ellipse in (â,b̂) space, e.g., the ellipse $\chi^2 = \chi^2_{min} + k^2$, are found by multiplying the term in curly brackets at the above figure by k.

Residuals and Goodness of Fit

Let us define the residual as $r_i = y_i - \hat{y}_i$. This is an analogue of the true error ϵ_i. Since ϵ_i is not accessible to us, we study r_i. If our model is correct, the distribution of the residuals should follow a certain form. Therefore, we look at the residuals in order to test the model. A plot of residuals can tell us if something is wrong with the fit or the data. In particular, a region of anomalously large residuals suggests that the model is not valid in that region of x. For example, if the errors ϵ_i are normal, then the residuals r_i should follow a normal distribution, each with its own x-dependent variance, derived from the above error hyperbolas. From the normal equations,

$$\sum r_i = \sum(y_i - \hat{y}_i) = \sum[y_i - \mathbf{H}(x_i)\hat{\theta}] = 0\,,$$

if there is a constant term in the model for y as a function of x and if the σ_i^2 are all identical, since, writing things out,

$$\hat{y} = \sum_j h_j(x)\hat{\theta}_j\,, \text{ and } \frac{\partial\chi^2}{\partial\theta_k} = -2\sum_i \frac{1}{\sigma_i^2}[y_i - \sum h_j(x_i)\theta_j]h_k(x_i)\ .$$

Now let $h_1(x_i) = 1$ (as in our straight-line example here) and find the χ^2 minimum:

$$\frac{\partial\chi^2}{\partial\theta_1} = 0 = \sum_i \frac{1}{\sigma_i^2}(y_i - \hat{y}_i) \times 1\ .$$

Therefore, if $\sigma_i^2 = \sigma^2$ for all i, $\sum r_i = 0$. Otherwise, $\sum \frac{r_i}{\sigma_i^2} = 0$ for diagonal $\mathbf{V}$. It is more complicated for general $\mathbf{V}$.

In matrix notation, we write

$$\vec{r} = \vec{y} - \hat{y} = \vec{y} - \mathbf{H}\,\hat{\theta} = \vec{y} - \mathbf{H}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{V}^{-1}\vec{y}$$

$$= (\mathbf{I} - \mathbf{R})\vec{y};\ \mathbf{R} = \mathbf{H}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{V}^{-1}\ ,$$

and therefore

$$\vec{r} - E(\vec{r}) = (\mathbf{I} - \mathbf{R})\,(\vec{y} - \mathbf{H}\,\vec{\theta}) = (\mathbf{I} - \mathbf{R})\vec{\epsilon}\,.$$

This follows from $E[(\mathbf{I} - \mathbf{R})\vec{y}] = (\mathbf{I} - \mathbf{R})E(\vec{y})$ and $E(\vec{y}) = \mathbf{H}\vec{\theta}$ if the model is correct. Of course, $\mathbf{RH}\vec{\theta} = \mathbf{H}\vec{\theta}$ and $E(\vec{r}) = 0$ (whether or not there is a constant term); the above expression enables us to casts what follows in terms of $\vec{\epsilon}$. Now,

$$V(\vec{r}) = E\{[\vec{r} - E(\vec{r})]^T[\vec{r} - E(\vec{r})]\} = (\mathbf{I} - \mathbf{R})^T\,E(\epsilon^T\epsilon)\,(\mathbf{I} - \mathbf{R})\,,$$

since the variance matrix of any vector $\vec{A}$ is $E\{[\vec{A} - E(\vec{A})]^T\,[\vec{A} - E(\vec{A})]\}$. If $\vec{A} - E(\vec{A}) = \mathbf{B}\vec{\epsilon}$, $\mathbf{B}$ a fixed, known, matrix, $V(\vec{A}) = \mathbf{B}^T E(\epsilon^T\epsilon)\mathbf{B}$.

Note that:

$$(\mathbf{I} - \mathbf{R})^T = \mathbf{I}^T - \mathbf{R}^T = \mathbf{I} - \mathbf{R}^T = \mathbf{I}^T - [\mathbf{H}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\,\mathbf{H}^T\mathbf{V}^{-1}]^T\ .$$

If $V(\vec{\epsilon}) = \mathbf{I}\sigma^2$, $\mathbf{R}^T = \mathbf{R} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\,\mathbf{H}^T$ and

$$\mathrm{V}(\vec{r}) = \mathrm{V}(\vec{y}) - \mathrm{COV}(y,\hat{y}) - \mathrm{COV}(y,\hat{y}) + \mathrm{V}(\hat{y})$$

$$= (\mathbf{I} - \mathbf{R})\, \mathbf{I}\sigma^2\, (\mathbf{I} - \mathbf{R}) = (\mathbf{I} - \mathbf{R})\, (\mathbf{I} - \mathbf{R})\, \sigma^2$$

$$= (\mathbf{I} - \mathbf{R} - \mathbf{R} + \mathbf{RR})\, \sigma^2 = (\mathbf{I} - \mathbf{R})\, \sigma^2\, [= \mathrm{V}(\vec{y}) - \mathrm{V}(\hat{y}) \text{ in this example}].$$

The last line follows since **R** is indempotent, i.e., **RR** = **R**:

$$\mathbf{RR} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\, \mathbf{H}^T\, \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\, \mathbf{H}^T = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\, \mathbf{H}^T = \mathbf{R}\text{, for diagonal } \mathbf{V}.$$

[In this example, it follows from $\mathrm{V}(r_i) = \mathrm{V}(y_i) - \mathrm{V}(\hat{y}_i)$ and the middle equation in the above set that $\mathrm{COV}(y_i,\hat{y}_i) = \mathrm{V}(\hat{y})$].

We note that $\mathbf{V}(\vec{r})$ has off-diagonal elements even though $\mathbf{V}(\vec{\epsilon})$ has none, i.e., there are correlations among the residuals. This is because a given point will tend to pull the curve towards it, thus affecting the other residuals. In the study of residuals, these correlations can usually be neglected, unless the number of points is not much more than the number of parameters being estimated.

We can take the error in one residual, r_i, as being $\sqrt{(1 - R_{ii})\sigma^2}$. When we speak of plotting the residuals, one should divide by this to arrive at standard normal variables. Many programs use just σ. In fact, this is probably okay most of the time, unless $k \cong n$.

If $\mathbf{V}(\epsilon)$ is not $= \sigma^2\mathbf{I}$ (i.e., the errors are not all expected to be identical, and/or there are correlations), then we must weight the residuals by a term expressing the error in the ϵ_i. We can always find a matrix **P** which looks like $\sqrt{\mathbf{V}}$ in the sense **PP** = **V**. This follows from the symmetric, positive definite nature of **V** [Basilevsky (1983)]. P will be symmetric, like V, i.e., $\mathbf{P}^T = \mathbf{P}$. If

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \cdot & \cdot & \cdot & 0 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 0 & \cdot & \cdot & \cdot & \sigma_n^2 \end{pmatrix}\text{, i.e., diagonal, } \mathbf{P} = \begin{pmatrix} \sigma_1 & \cdot & \cdot & \cdot & 0 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 0 & \cdot & \cdot & \cdot & \sigma_n \end{pmatrix},$$

which expresses the error of each ϵ_i. Then we should study $\vec{r}' = \mathbf{P}^{-1}\vec{r}$, [i.e., $r_i' = (\mathrm{P}^{-1}r)_i$, at a given x], divided by *its* error. For diagonal **P**, $r_i' = r_i/\sigma_i$. The variance of r' does not depend upon σ, and $r_i'/\sqrt{\mathrm{V}(r_i')}$ will look very similar to $r_i/\sqrt{\mathrm{V}(r_i)}$. Let us write:

$$\vec{r}' = \mathbf{P}^{-1}[\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\, \mathbf{H}^T\mathbf{V}^{-1}]\vec{y}$$

$$\equiv \mathbf{P}^{-1}[\mathbf{I} - \mathbf{R}]\vec{y}\text{; where, as above, } \mathbf{R} = \mathbf{H}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\, \mathbf{H}^T\, \mathbf{V}^{-1}\text{. Then}$$

$$\vec{r}' - \mathrm{E}(\vec{r}') = \mathbf{P}^{-1}\, [\mathbf{I} - \mathbf{R}]\, [\vec{y} - \mathbf{H}\,\vec{\theta}]$$

$$= \mathbf{P}^{-1}\, [\mathbf{I} - \mathbf{R}]\vec{\epsilon}\,.$$

Therefore

$$\mathrm{V}(\vec{r}') = \mathbf{P}^{-1}\, [\mathbf{I} - \mathbf{R}]^T\, \mathrm{E}(\epsilon^T\epsilon)\, [\mathbf{I} - \mathbf{R}]\, \mathbf{P}^{-1}$$

$$= \mathbf{P}^{-1}\, [\mathbf{I} - \mathbf{R}]^T\, \mathbf{V}[\mathbf{I} - \mathbf{R}]\, \mathbf{P}^{-1}$$

$$= \mathbf{P}^{-1}\, [\mathbf{I} - \mathbf{R}]^T\, \mathbf{PP}[\mathbf{I} - \mathbf{R}]\, \mathbf{P}^{-1}\,.$$

If $\mathbf{V}$ is diagonal, $\mathbf{R} = \mathbf{H}(\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H})^{-1}\,\mathbf{H}^T\,\mathbf{V}^{-1} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\,\mathbf{H}^T$, and

$\mathbf{V}(\vec{r}') = [\mathbf{I} - \mathbf{R}]^T\,[\mathbf{I} - \mathbf{R}]$ (not a function of σ or of $\mathbf{V}$, since $r'_i = r_i/\sigma_i$)

$= [\mathbf{I} - \mathbf{R}]$ since, as before for this case $\mathbf{R}^T = \mathbf{R}$ and $\mathbf{RR} = \mathbf{R}$.

Thus we should study, for diagonal $\mathbf{V}$ (independent measurements),

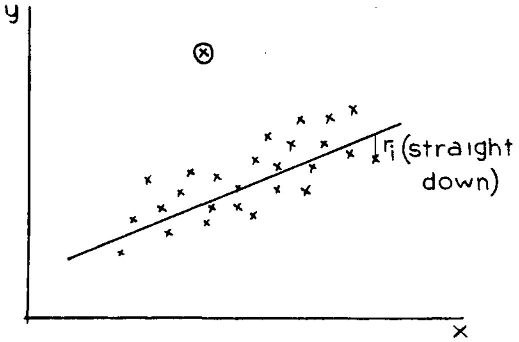
$$r'_i/\sqrt{V(r'_i)} = r_i/\sigma_i\sqrt{V(r'_i)}$$

$$= r_i/\sigma_i\sqrt{1 - R_{ii}} = \frac{r_i}{\sqrt{V(y_i) - V(\hat{y}_i)}}$$

to look for goodness of fit i.e., to see if there are any areas of significant deviation of the model from the data. It can be shown that if the ϵ_i are normal: $N(\epsilon_i\,;\,0,\,\sqrt{V_{ii}})$, then the $(\mathbf{P}^{-1}\,\epsilon)_i$ are also normal: $N[(\mathbf{P}^{-1}\,\epsilon)_i\,;\,0,\,1]$ (recall that the σ's are included in the $\mathbf{P}$ term). Then the $\vec{r}'_i$ are normal (multivariate) with variance given above.

Outliers

One of the things a study of residuals can do is detect the presence of *outliers.*



We should examine the circled point for possible rejection as bad data. The only really safe cause for rejection is if we can detect some malfunction in the apparatus when the point was taken. Otherwise, the outlier may just arise from the long tails we have talked about. Such a point can be deadly, because in L.S. analysis the important quantity is the distance squared.

The Gauss-Markov Theorem tells us that a L.S. fit is optimal even in the presence of outliers arising from the famous long tails appended to a Gaussian characteristic of many measuring processes, provided these long tails satisfy the assumptions of being unbiased. The problem arises because outliers are often badly biased.

The study of outlier rejection is not very mature. Frequently we reject if (a) the point is very unlikely according to our model of the distribution of the measurements, and (b) if there is a large gap in the distribution of residuals between the location of the one point and the rest of the data. Point (b) runs into trouble if there exists more than one outlier.

Goodness of Fit

Once we have studied our data for the presence of outliers, and have decided what to do about any we find, we have done much of what we can do from a statistical point of view to ensure we are fitting to reasonable data. We now wish to study goodness of fit to see if the model is a reasonable description of the data.

If we reject outliers, our fit procedure corresponds to fitting to a truncated ϵ_i distribution. If there is no large gap, we have some difficulties identifying outliers because we need to know the parameters of the fit, to give us the distribution of the residuals, before we can truncate. One approach to this problem is to make a first guess, truncate, fit, then re-examine the residuals and do a new truncation and a re-fit, etc. hoping the procedure converges. Because of the Gauss-Markov Theorem, L.S. works even with a truncated normal, so this is okay. One must adjust the variance, since **V** of the untruncated normal no longer represents the true variance of $\vec{y}$. The true variance of $\hat{y}$ will have smaller diagonal elements than in **V**. This does not mean that $\mathbf{V}(\hat{\theta})$ is reduced, because the loss of information from discarding some points (which means we have to be prepared to reject some perfectly valid points hidden among our outliers) will more than compensate for the reduction of **V**. The derivative formulation will give the correct errors.

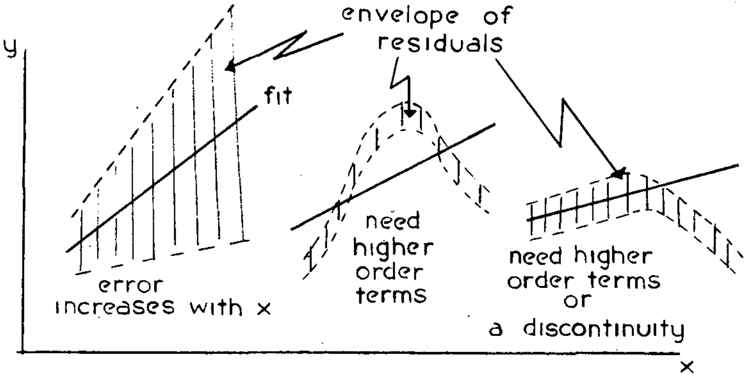
In a computer program, one approach frequently taken is to (a) examine the overall χ^2, (b) if the χ^2 is improbably large, reject the point with largest r_i^2, (c) re-fit with the point gone and test the rejected point again. The test statistic to see if the point should have been rejected can be taken to be the *new* residual squared for the point, i.e., distance squared from the new line. Compare this with $\chi^2(1)$ and see if it was unlikely. This works okay if $n >> k$. If we test just on r_i^2 from the overall fit to all the points, any outliers will affect the fit in their favor, because they are so important; therefore we might never reject a point. In any case, the overall χ^2 has contributions from every data point, and an unusually large contribution from one particular point may not be noticed if we don't look at the residuals; we might conclude that we should accept the fit even though one bad data point has affected the fitted parameters appreciably.

The danger with rejection of outliers is that we may be throwing away a real effect, e.g., a resonance. Important discoveries have come from careful study of outliers.

If our model for the measurements is that they should be normal, then we can look at the overall χ^2. Or, perhaps better if we are doing a lot of fits, to plot $P(\chi^2) = \int_{\chi^2_{obs}}^{\infty} \chi^2(n - k)d\chi^2$ for all the fits. This should be uniform. Often there is a peak at low $P(\chi^2)$ because the measurements are "heavy-tailed" normals rather than true normals. When this occurs, we should bear in mind that the least-squares fit will still be unbiased provided only that all the measurements, including especially those in the heavy tails, are unbiased. The point at which we should become suspicious of that situation will depend on the experiment. The point at which we should reject a fit, biased or not, will also depend on the experiment, including our considerations such things as the distribution of $P(\chi^2)$ which we find and the possible presence of competing fits with much higher $P(\chi^2)$.

There are other approaches to estimation designed to alleviate this sort of problem. "Robust" statistics are insensitive to heavy tails. Although not usually as precise if there are no heavy tails, robust estimators can be much better than those we are discussing if, say, the normal distribution we assume for the measurement errors is "contaminated" at the 10% level by a much broader normal. Rey (1983) is a good reference for a discussion of this.

By plotting the residuals, or, better, the residuals divided by their error, we can also look for patterns, e.g.,



These types of pattern suggest a failure of the model to describe the data. (Remember, we can never prove a model to be correct; we can only prove that a competing model is highly unlikely to be correct).

The more residuals, the more powerful our test of the model.

Note: To optimize our test of the model against arbitrary alternatives, we need points, and therefore residuals, more or less evenly distributed along the curve (obviously, if we are suspicious of a particular region of x, we obtain a better test if we concentrate points in that region). However, to optimize our measurement precision (i.e., to minimize the variance of θ) we want to cluster all our measurements at the extreme two points for two-dimensional θ, at the properly-chosen three points for three-dimensional θ, etc. This provides us with little or no power to evaluate the model. Clustering points at a particular x does, however, allow us to test the measurement process independent of the model, since we can compare the distribution of y_i with expectations. That is, at a fixed x_i, the distribution of y_i cannot depend upon the model assumptions; it can only depend upon the measuring process, which we believe we understand. This sometimes reveals an unsuspected dependence on some variable neglected in the model, e.g., temperature of the sample, or whatever.

Unknown σ^2

If σ^2 is unknown, then we must estimate it from the data. If there are n data points, and we were estimating k parameters, then an unbiased guess for the value of σ^2, assumed to be the same for all i, is

$$S^2 = \sum_i \frac{(r_i - \overline{r})^2}{(n - k)} = \frac{\sum r_i^2}{n - k} \,;$$

the last equality holds if $\overline{r} = 0$. Then we plot $r_i/[(1 - R_{ii})\, S^2]^{1/2}$ when we study residuals.

Then, as well, our variances acquire a dependence on the $\chi^2 = (n - k)S^2$:

$$\hat{V}(\hat{b}) = \frac{S^2}{\sum(x_i - \overline{x})^2}$$

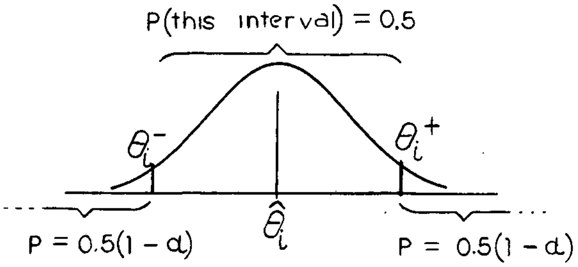
$$\hat{V}(\hat{a}) = \frac{S^2 \sum x_i^2}{n \sum(x_i - \overline{x})^2} \text{ , for a straight-line fit.}$$

This doesn't change the L.S. estimates of $\hat{a}$ and $\hat{b}$, since $V(\epsilon)$ could be expressed as a weight matrix $\times$ σ^2, and σ^2 divides out in $\hat{\theta}$.

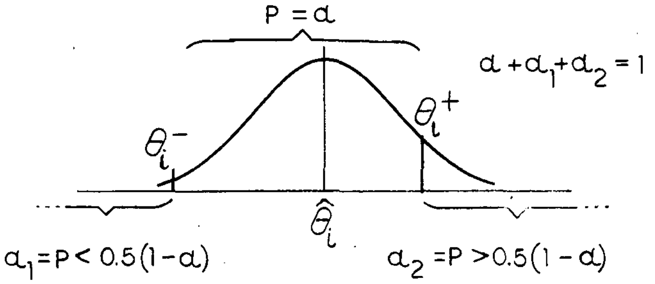
Up to now, we have not discussed the relationship of the variance to a probability for the true θ to lie in a certain interval, except for the Gaussian. In that case, the true value lies within 1.0 σ, where $\sigma = \sqrt{V}$, 68% of the time, within 2.0 σ 95% of the time, etc (Chapter 3). Now we have a more complex case, and it is appropriate to discuss *confidence intervals.*

Confidence Intervals

Let us set up confidence intervals for the θ_i. We pick limits θ_i^+ and θ_i^- such that $P(\theta_i^- \leq \theta_i \leq \theta_i^+) = 100\alpha\%$, α a pre-selected number < 1. "α" is a proability, called the confidence level, which expresses our estimate of the probability that the interval we specify actually includes the true value. We can often do this only approximately. If the θ_i are Gaussian (which requires ϵ_i Gaussian with known σ_i^2), we usually choose a symmetric interval:



We could choose an unsymmetric interval:



We must include the error in S^2 when we want to estimate a confidence interval for our parameters, for example $\hat{a}$ and $\hat{b}$. That is, the probability that the line can vary by a certain amount must take account of the variation of our *error estimates* as we vary the parameters of the line. This introduces a broadening in the distribution for $\hat{a}$ and $\hat{b}$, which would otherwise be Gaussian. This broadening is not very substantial if $n - k$ is large.

The limits of the confidence interval, θ_i^- and θ_i^+, are obtained by using the cumulative distribution function (c.d.f.) to evaluate the probabilities. If the c.d.f. for $\hat{\theta}$ is F, and α, α_1, and α_2 are as defined on the figure,

$$F(\theta_i^+) = \alpha_1 + \alpha \text{ and } F(\theta_i^-) = \alpha_1 \ .$$

If we choose a symmetric interval, then $\alpha_1 = \frac{1}{2}(1 - \alpha)$ and θ_i^+ is chosen from $F(\theta_i^+)$; similarly for θ_i^-. There are not many general guidelines for selecting from the infinite possible choices of confidence intervals, but an unsymmetric interval is longer than a symmetric interval if the underlying distribution is symmetric and peaks at $\hat{\theta}_i$. Usually, the shortest interval for a given confidence level α is the interval of choice.

To set up confidence intervals for the unknown σ^2 case, we will need:

Student's t Distribution

If the Gaussian model (or whatever model we are using for the distribution of the points) is correct, note that we will be wrong about θ a fraction $1 - \alpha$ of the time, i.e., the α confidence level interval we specify will not contain θ_{true} that specified fraction of the time. It can be shown that we can set up $100\alpha\%$ ($\alpha < 1$) confidence intervals (e.g., 68%, 95%, etc.) for θ_i, for Gaussian errors of unknown (but equal) size in the y_i, from

$$\hat{\theta}_i \pm T[\frac{1}{2}(1 + \alpha); n - k] \times \sqrt{V(\theta_i)} \quad \text{, e.g.,}$$

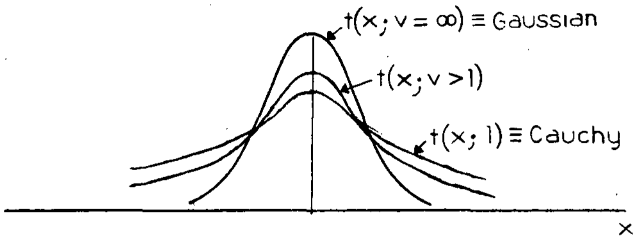
$$\hat{a} \pm T[\frac{1}{2}(1.68); n - 2] \times S \times \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}$$

for the intercept of the straight line at a 68% confidence level.

and something similar for $\hat{y}$. That is, one standard deviation in θ_i, $\sqrt{V(\theta_i)}$, does not in this case represent 68% of the probability; it represents somewhat less. In the limit $n - k \gg 1$, $T[\frac{1}{2}(1.68), n - k] \rightarrow 1$. T is related in a manner we shall show to the c.d.f. of t, which is a p.d.f. we haven't yet studied, called Student's t distribution:

$$t(x;\, v) = \frac{(\frac{v-1}{2})\,!}{(\frac{v-2}{2})\,!\,\sqrt{\pi v}} \left[1 + \frac{x^2}{v}\right]^{-\frac{v+1}{2}} \quad ;$$

v an integer parameter = n − k.



The t is the distribution of a variable

$$x = \frac{Z}{\sqrt{\chi^2(v)/v}}$$

where Z is a standardized normal [N(z;0,1)], and χ^2 is an independent χ^2 r.v. of v degrees of freedom. As v increases (i.e., we get more data points), the distribution becomes very nearly the same as a normal because the estimate S^2 for σ^2 becomes more accurate.

The function T is not the c.d.f. of t, but is derived from it. The value of T is the *limit of integration* of the p.d.f. up to a specified probability, i.e., the argument of the c.d.f:

$T[\frac{1}{2}(1 + \alpha), n - k]$ is defined as y such that

$$\int_{-\infty}^{y} t\,(x;\, n - k)\, dx = \frac{1}{2}(1 + \alpha)\,.$$

Note that T is a dimensionless number, since x is; the scale is set by $\sigma = \sqrt{V}$. If $n \gg k$ or if σ^2 is fixed in advance, T becomes the argument of the error function: T is chosen such that $\mathrm{erf}(T) = \frac{1}{2}(1 + \alpha)$.

For *example*, a 95% confidence level interval:

30 points, k = 2, 100α = 95% → T(.975, 28) = 2.04
10 points, k = 2, 100α = 95% → T(.975, 8) = 2.23

5 points, k = 2, 100α = 95% → T(.975, 3) = 2.78
3 points, k = 2, 100α = 95% → T(.975, 1) = 12.71

whereas, if we just assumed that our θ_i was Gaussian, a 95% C.L. would give exactly 2.0, i.e., two standard deviations, independent of n (since we don't then need the data to estimate σ^2). If n is only k + 1, the necessity of estimating σ^2 extends this interval to 12.71. That is, the probability that the symmetric interval from $\hat{\theta} - 12.71\sqrt{V}$ to $\hat{\theta} + 12.71\sqrt{V}$ contains the true value is 95%.

Up to now, we have assumed that the predictor variable x is determined without error. If we relax that assumption, things get a bit more complicated.

Errors in Both x and y: Point Fit

Let us now suppose we have a set of n independent measurements $\vec{z}_i = (x_i, y_i)$, where both x and y have errors, with variance matrix $\mathbf{V}_i(2 \times 2)$ for each point. Let us discuss the simple model that these are all measurements of the same quantity, i.e., the model is not a curve or a line but a single point. Then the true value of that quantity is

$$\vec{z}^* = (x^*, y^*), \text{ and}$$

$$\chi^2 = \sum_i (\vec{z}_i - \vec{z}^*)^T \mathbf{V}_i^{-1} (\vec{z}_i - \vec{z}^*),$$

which is a χ^2, of course, only if the $\vec{z}_i$ are bivariate normals. This is a generalization of our previous results. If we regard $\vec{z}^*$ as an unknown variable which we must evaluate from the data,

$$\frac{\partial \chi^2}{\partial \vec{z}^*} = 2 \sum_i (\vec{z}_i - \vec{z}^*)^T \mathbf{V}_i^{-1} = 0 \text{ at } \vec{z}^* = \hat{z} \text{ for the L.S. solution.}$$

We assume that V_i is a sufficiently slowly-varying function of x and y that we can neglect its variation in all derivatives compared with the other terms.

Each measurement has two components. Now we write, solving the above,

$$\hat{z} = (\sum \mathbf{V}_i^{-1} \vec{z}_i)(\sum \mathbf{V}_i^{-1})^{-1},$$

which is clearly a generalization of our earlier result in one dimension:

$$\hat{\mu} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}.$$ The variance in $\hat{z}$ is also analogous:

$$\begin{aligned}\mathbf{V}(\hat{z}) &= (\sum \mathbf{V}_i^{-1})^{-1}[\sum \mathbf{V}_i^{-1}\, \mathbf{V}(\vec{z}_i)\mathbf{V}_i^{-1}](\sum \mathbf{V}_i^{-1})^{-1} \\ &= (\sum \mathbf{V}_i^{-1})^{-1} (\sum \mathbf{V}_i^{-1}) (\sum \mathbf{V}_i^{-1})^{-1} \\ &= (\sum \mathbf{V}_i^{-1})^{-1}\end{aligned}$$

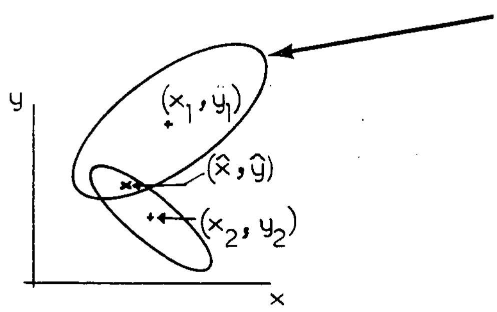
This follows from propagation of errors, since $\mathbf{V}(\hat{z})$ will be a sum of terms from each independent $\vec{z}_i$ and each term will be $\left(\frac{\partial \hat{z}}{\partial z_i}\right)^T \mathbf{V}(\vec{z}_i) \frac{\partial \hat{z}}{\partial z_i}$, where $\frac{\partial \hat{z}}{\partial z_i}$ is a 2 × 2 symmetric matrix equal to

$(V_i^{-1})(\Sigma V_i^{-1})^{-1}$.

Obviously, there was no requirement that all the z_i were measured by a single experiment. We might have several experiments which measure the same two-component vector quantity; this shows us how to combine these experiments.

The same equations hold, of course, in any number of dimensions, not just two.

Graphically, we have something like this

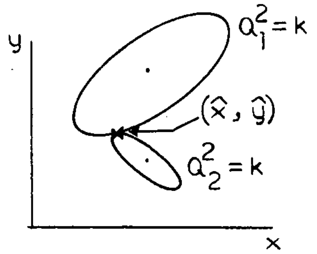


Q^2 fixed, e.g., $Q^2 = 1$

[Q^2 = value of χ^2 from one data point; contours of constant Q^2 shown]

Ellipses if the errors on (x_i, y_i) are Gaussian. If these errors are not Gaussian, the above error formula is the leading term in the propagation of errors expansion. We may still get ellipses in that case, also, but the probability enclosed by the ellipses would differ from the Gaussian case.

For just two measurements (n = 2), the solution will occur at the point where the *equivalent* ellipses (i.e., same value of Q^2) first touch:

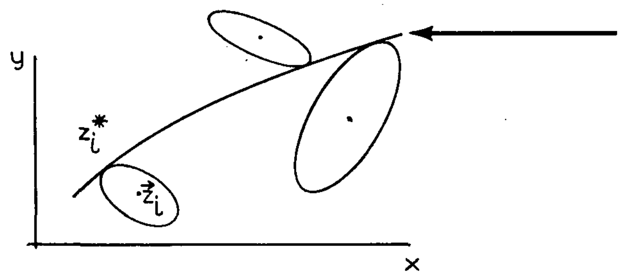


$\chi^2 = Q_1^2 + Q_2^2$;
minimum at $\chi^2 = 2k$

If we consider the x and y residuals separately (each one is in a different dimension), in each case we have n = 2, k = 1. Then, we will have $r_{x_1}/\sigma_{r_{x_1}} = r_{x_2}/\sigma_{r_{x_2}}$ and similarly for y. This is a special case: $n - k = 1$. This is called a one *constraint* fit because we have one measurement more than the minimum necessary (which is one) to determine the parameter of interest, $\hat{x}$ or $\hat{y}$. In a one constraint fit, each measurement contributes equally to the χ^2.

Errors in Both x and y: Line Fit

Now, let's suppose we have to determine a line instead of a point. Each measured point will contribute to the χ^2 according to the following picture:



y = f(x), a candidate for the best fit to the points

If the model is correct, each point of contact is our best guess at the true value which we were trying to measure when we actually obtained $\vec{z}_i$.

The generalized distance squared to the line is $d_i^2 = (\vec{z}_i - \vec{z}_i^*)^T V_i^{-1} (\vec{z}_i - \vec{z}_i^*)$ where $\vec{z}_i = (x_i,y_i)$ and $\vec{z}_i^*$ is the point on the line. The χ^2 is

$$\chi^2 = \sum d_i^2 = \sum(\vec{z}_i - \vec{z}_i^*)^T V_i^{-1} (\vec{z}_i - \vec{z}_i^*) \ ,$$

if the points are uncorrelated (V_i is two-dimensional, as above, representing the two-dimensional error ellipse about $\vec{z}_i$). So far, we haven't said anything about $\hat{z}$.

We must satisfy the condition that $\hat{y} = \hat{H}^T \hat{\theta}$, where now H can vary too, i.e., $\hat{H} = H(\hat{x})$. This is still the linear model. But whereas before we wrote y as a vector, one element for each x_i, let us now write y as a one-element scalar function, a function of x, and

$$H(x) = \begin{pmatrix} h_1(x) \\ h_2(x) \\ \bullet \\ \bullet \\ \bullet \\ h_k(x) \end{pmatrix}$$

i.e., as a vector function of x. Then $\hat{y}(x) = H^T(\hat{x})\hat{\theta}$.

How many unknowns do we have? We need k θ's and, in some cases, we also want to know the $(\hat{x}_i,\hat{y}_i) = [\hat{x}_i, H^T(\hat{x}_i)\ \hat{\theta}]$, which gives us just 2n more unknowns. Then we have 2n + k unknowns, altogether. The best approach to solve for these unknowns is to use Lagrange's undetermined multipliers. We wish to minimize χ^2 subject to the constraint $\hat{y} = H^T(\hat{x})\hat{\theta}$. Thus, we set up n unknown multipliers λ_i, one for each z_i. These are n new unknowns, but they come with their own n equations. We minimize

$$\chi^2 = \sum_i [(\vec{z}_i - \vec{z}_i^*)^T V_i^{-1}(\vec{z}_i - \vec{z}_i^*) + 2\,\lambda_i(y_i^* - H^T(x_i^*)\,\theta)]$$

↑

for convenience later

with respect to the k θ's and the n λ's. (We use the "*" to represent some arbitrary value. The "*" is replaced by a "$\wedge$" at minimum. The measured point is $\vec{z}_i$.) The first term is the contribution due to the fact that we insist on a value z* which is different from the measured value (except in unusual cases) when the constraints are satisfied. The second term allows us to incorporate the constraints of the model into the χ^2 minimization by requiring the χ^2 to be minimum in the new λ_i dimensions.

Okay, now

$$\frac{\partial \chi^2}{\partial \theta_\ell} = \sum_i [-2V_i^{-1}(\vec{z}_i - \hat{z}_i)\,\frac{\partial \hat{z}_i}{\partial \theta_\ell} - 2\,\lambda_i\,H_\ell\,(\hat{x}_i)] = 0\,, \quad \ell = 1\,, \ldots, k\,;$$

$$\frac{\partial \chi^2}{\partial \lambda_i} = 2\,(\hat{y}_i - H^T\,(\hat{x}_i)\,\hat{\theta}) = 0\,,\ i = 1, \ldots, n\,.$$

The second equation just recovers the constraints, i.e., that the model be satisfied. We see that the arbitrary factor of 2 does not affect the χ^2 at the solution, since the term being multiplied becomes zero. In these equations, we will need:

$$\frac{\partial \hat{z}_i}{\partial \theta_\ell} = \frac{\partial}{\partial \theta_\ell}\begin{pmatrix}\hat{x}_i \\ \hat{y}_i\end{pmatrix} = \frac{\partial}{\partial \theta_\ell}\begin{pmatrix}\hat{x}_i \\ H^T(\hat{x}_i)\hat{\theta}\end{pmatrix} = \begin{pmatrix}\hat{x}_i \\ H_\ell^T(\hat{x}_i)\end{pmatrix}\ .$$

We have k unknown θ's
n unknown x_i's.
n unknown y_i's.
n unknown λ_i's.

So far we have n + k equations.

Finally,

$$\frac{\partial \chi^2}{\partial x^*} = 0 \text{ at } x^* = \hat{x}\,, \text{ (n equations) ; and}$$

$$\frac{\partial \chi^2}{\partial y^*} = 0 \text{ at } y^* = \hat{y}\,, \text{ (n equations) .}$$

These two conditions yield

$$-\ 2[\mathbf{V}_i^{-1}(z_i - \hat{z}_i)]_1 - 2\lambda_i\,H'^T(\hat{x}_i)\hat{\theta} = 0 \quad (H' = \frac{\partial H}{\partial x})\,,\ \text{and}$$

$$-\ 2[\mathbf{V}_i^{-1}(z_i - \hat{z}_i)]_2 - 2\lambda_i = 0\,.$$

The symbols $[\ ]_1$ and $[\ ]_2$ refer to the first and second components, respectively. These are 2n equations which just express the fact that for a given line, i.e., a given $\hat{\theta}$, $(\hat{x}_i,\hat{y}_i)$ is that point at which the covariance ellipse, $\chi_i^2 = (z_i - \hat{z}_i)^T\mathbf{V}_i^{-1}\,(z_i - \hat{z}_i)$ is, just tangent to the curve. We can solve for λ_i easily enough:

$$\lambda_i = [\mathbf{V}_i^{-1} (z_i - \hat{z}_i)]_2 \; .$$

Then we can plug in $\hat{y}_i = H^T(\hat{x}_i)\hat{\theta}$ and solve the other equation immediately above for $\hat{x}_i$ as a function of $\hat{\theta}$. If there are any H functions which are non-linear in x (i.e., something other than "1" or "x"), these equations are a set of coupled, non-linear equations. Having solved for $\hat{x}$ as a function of $\hat{\theta}$, and λ_i as above, we now plug these expressions into the remaining equation (for $\partial\chi^2/\partial\theta_\ell$) and solve for $\hat{\theta}$. This is obviously very complex, and we shall not cover it here in full generality. Most experimenters will solve the problem numerically. But there are some very useful cases we can cover in more detail.

Let us suppose we have an unbiased set of measurements of different quantities $(\theta_1,\theta_2,\ldots,\theta_n)$, which we suppose to be the quantities of direct interest. That is, we are not interested in finding a curve that fits them and then publishing the parameters of the curve. We will fit a curve, but it's the values of θ_i that lie on the curve that we are after. The θ_i's play the role of the x_i or y_i in the above discussion; we use the notation θ_i so that θ continues to represent the quantities of physical interest. $\vec{\theta}$ may be a vector of arbitrary dimensionality with components θ_i. Let us suppose that we have a set of k constraint equations which are linear relationships among the θ's which must be satisfied:

$$\underset{\substack{\uparrow \\ (k\times n)}}{\mathbf{B}} \quad \underset{\substack{\uparrow \\ (n\times 1)}}{\vec{\theta}} \; = \; \underset{\substack{\uparrow \\ (k\times 1)}}{b} \quad ; \text{ k constraints for n } \theta\text{'s.}$$

Example

3-dimensions (n = 3).

$$k = 1: \quad (B_1 \; B_2 \; B_3) \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = b \; .$$

Hence $B_1\theta_1 + B_2\theta_2 + B_3\theta_3 = b$. This is the general equation of a plane. The solution is constrained to lie upon this 2-dimensional surface.

$$k = 2: \quad \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ & & \\ B_{21} & B_{22} & B_{23} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ \\ b_2 \end{pmatrix} \; .$$

Hence $B_{11}\theta_1 + B_{12}\theta_2 + B_{13}\theta_3 = b_1$ and simultaneously $B_{21}\theta_1 + B_{22}\theta_2 + B_{23}\theta_3 = b_2$. For the constraints to admit of a solution, it must lie upon the one-dimensional line defined by the intercept of these two planes.

k = 3: Clearly if the solution space is not empty, it consists of a single point which is the point of intersection of the three planes defined by the constraint equation. No fitting is necessary in this case; there is no freedom in chosing the solution to satisfy any statistical considerations.

For formulas relevant to planes which intersect, see, e.g., Bartsch (1974), Chapter 4.

Example.

The Table. The straight, parallel legs of an ordinary table have lengths θ_i. Let the table have n legs. Then for the legs to have lengths all equal to b, the constraint equation is

$$\mathbf{I}\vec{\theta} = \vec{b} \ ,$$

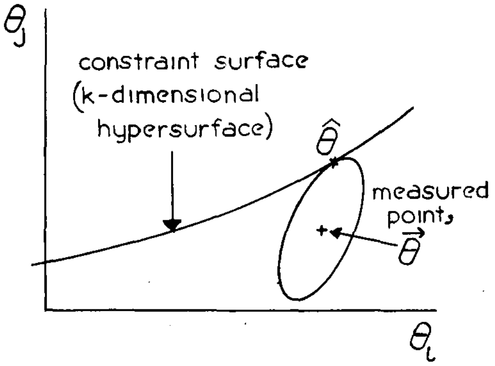
where $\vec{b}^T = (b, b, ..., b)$. If we wish to specify that the legs must be equal in length, but we don't care what that length is, then $\theta_1 = \theta_2 = \theta_3$ for n = 3 legs. This suggests three planes $\theta_1 - \theta_2 = 0$, $\theta_2 - \theta_3 = 0$, $\theta_1 - \theta_3 = 0$, of which any two will suffice to define a straight line by their intersection; the third is trivially shown to be redundant. Then, for example,

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = 0$$

is the constraint equation.

No construction technique can cause these constraints to be satisfied exactly. We depend upon minute deformations in the table-top, the legs, and the floor to compensate. But if the θ_i's are too far from satisfying the conditions, the data are inconsistent with the model—the table is not level, or it rocks.

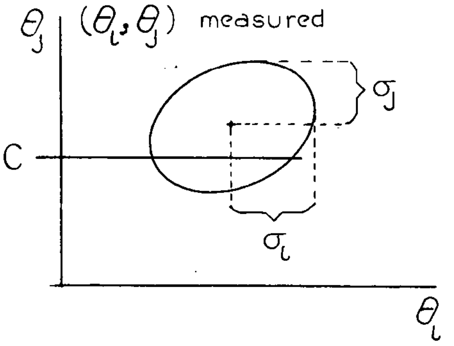
For a physics example, the θ's might be the three-momenta of a set of tracks, and the constraint equations are just the equations of momentum conservation. (If we add energy conservation, it becomes non-linear, to be discussed later.) If $\vec{\theta}$ is viewed as a point in n-space, then the constraint equations restrict us to a hypersurface of n $-$ k dimensions in which the true point must lie. In the presence of non-zero measurement error, $\vec{\theta}$ as measured won't satisfy the constraint equations, in general. We want to find $\hat{\theta}$ satisfying these equations, such that the $\hat{\theta}$'s are as "close" to the $\vec{\theta}$'s as possible, where we define "closeness" in a Least-Squares sense. Then we can use the χ^2 we obtain to decide if the data are close enough that we can say the model is consistent with the data.



2 out of n dimensions displayed

The relationship with the preceding discussion is that there we had n two-dimensional points $\vec{z}_i$; here we have one n-dimensional point $\vec{\theta}$. There we had a functional form $y = H^T\vec{\theta}$ whose vector of coefficients $\vec{\theta}$ was to be determined from the measured quantities y_i and $H^T(x_i)$. Here we have a pre-determined functional form $\mathbf{B}\vec{\theta} = \mathrm{b}$ and we don't need to find anything more than the value of $\vec{\theta}$ satisfying that equation which lies closest to the measured $\vec{\theta}$, in the L.S. sense.

If the model is correct, it is only random measurement error that causes $\vec{\theta}$ to be away from the constraint surface. If we adjust the $\vec{\theta}$'s until we satisfy the constraints, with a resultant improved set of estimates $\hat{\theta}$ for the true $\vec{\theta}$'s, we have a physically consistent set of measurements with errors reduced by the condition that the constraints be satisfied. For example,



Bivariate Normal θ_i and θ_j

If the constraint equation is $\theta_j = C =$ constant, then θ_i is described as a Gaussian with variance $V(\theta_i|\theta_j) = \sigma_i^2(1 - \rho^2) \leq \sigma_i^2$, as we have seen before.

We will use Lagrange's multipliers. Let us define $\vec{\theta}$ as being the originally-measured values. Then, at some other value θ^*,

$$\chi^2(\theta^*) = (\vec{\theta} - \theta^*)^T\, \mathbf{V}^{-1}\, (\vec{\theta} - \theta^*) + 2\lambda^T\, (\mathbf{B}\theta^* - \mathrm{b})\,. \text{ (There are k } \lambda\text{'s)}$$

$\mathbf{V}$ is the full covariance matrix for $\vec{\theta}$, including any correlations. The normal equations become

$$\frac{\partial \chi^2}{\partial \theta^*} = -\,2\,\mathbf{V}^{-1}(\vec{\theta} - \hat{\theta}) + 2\,\mathbf{B}^T\lambda = 0 \text{ at } \theta^* = \hat{\theta}\,; \text{ therefore}$$

$$\mathbf{B}^T\lambda = \mathbf{V}^{-1}\,(\vec{\theta} - \hat{\theta})\,.$$

$$\frac{\partial \chi^2}{\partial \lambda} = 2(\mathbf{B}\hat{\theta} - \mathrm{b}) = 0 \text{ at } \theta^* = \hat{\theta} \text{ recovers the constraints, as before.}$$

$$\vec{\theta} - \hat{\theta} = \mathbf{VB}^T\lambda\,;\ \mathbf{B}\vec{\theta} - \mathbf{B}\hat{\theta} = \mathbf{BVB}^T\lambda\,;\ \mathbf{B}\vec{\theta} - \mathrm{b} = \mathbf{BVB}^T\lambda$$

(involving λ as the only unknown) .

Therefore $\lambda = (\mathbf{BVB}^T)^{-1}\,(\mathbf{B}\vec{\theta} - \mathrm{b})$, and therefore $\mathbf{V}^{-1}\,(\vec{\theta} - \hat{\theta}) = \mathbf{B}^T\lambda$ gives us

$$\hat{\theta} = [\mathbf{I} - \mathbf{V}\,\mathbf{B}^T\,(\mathbf{BVB}^T)^{-1}\,\mathbf{B}]\,\vec{\theta} + \mathbf{V}\,\mathbf{B}^T\,(\mathbf{BVB}^T)^{-1}\,\mathrm{b}\,.$$

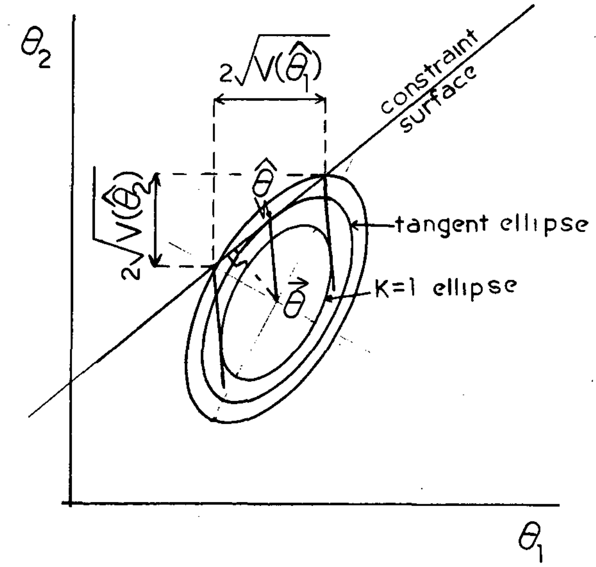
This gives us the correction to the measured $\vec{\theta}$ to satisfy the constraints at minimum χ^2. Let us recall here that $\mathbf{B}$ is not square and has no inverse unless $k = n$. Otherwise we would have $\mathbf{V}\,\mathbf{B}^T(\mathbf{BVB}^T)^{-1}\mathbf{B} = \mathbf{B}^{-1}(\mathbf{BVB}^T)(\mathbf{BVB}^T)^{-1}\mathbf{B} = \mathbf{I}$, and $\hat{\theta} = \mathbf{B}^{-1}\mathrm{b}$. If $k = n$ and $\mathbf{B}$ is non-singular, the problem is completely determined without error. The measurements are irrelevant, except to test goodness of fit. For $k < n$,

$$E(\hat{\theta}) = E(\mathbf{I}\vec{\theta}) - \mathbf{V}\,\mathbf{B}^T\,(\mathbf{BVB}^T)^{-1}E(\mathbf{B}\vec{\theta} - \mathrm{b})$$

$$= \vec{\theta}_{\text{true}}\,, \text{ since } \mathbf{B}E(\vec{\theta}) = \mathrm{b}\,; \text{ therefore } \hat{\theta} \text{ is unbiased.}$$

It is easily demonstrated that the form $\mathbf{A} = \mathbf{I} - \mathbf{V}\mathbf{B}^{\mathbf{T}}(\mathbf{B}\mathbf{V}\mathbf{B}^{\mathbf{T}})^{-1}\mathbf{B}$ is idempotent (i.e., the matrix product $\mathbf{A} \cdot \mathbf{A} = \mathbf{A}$), which is a necessary and sufficient condition for **A** to be a projection operator [Basilevsky (1983), Section 4.8]. **A** projects $\vec{\theta}$ onto the hypersurface represented by the constraint equation. If **V** commutes with the symmetric factor $\mathbf{B}^{\mathbf{T}}(\mathbf{B}\mathbf{V}\mathbf{B}^{\mathbf{T}})^{-1}\mathbf{B}$, which is the case if (among other possibilities) **V** is a multiple of the indentity matrix, $\mathbf{V} = \sigma^2\mathbf{I}$, i.e., the measurements are independent and identically distributed, then **A** is symmetric and the projection is orthogonal.

The situation is illustrated by the figure below. In two dimensions the linear constraint reduces to a straight line upon which the solution must lie. An orthogonal projection of the measured point is shown by the dashed line. But the error ellipses, contours of constant χ^2, are inclined at an angle to the line. A better χ^2 is obtained by the apparently longer non-orthogonal projection to $\hat{\theta}$, which lies at the point at which the smallest possible ellipse (smallest χ^2) touching the line is just tangent to it. In the case of independent, identically distributed measurements θ_i, these ellipses become circles with no preferred direction and the orthogonal projection is best. If the ellipses happen to lie with one axis orthogonal to the constant line, an orthogonal projection will also be best.



From propagation of errors we get

$$\mathbf{V}(\hat{\theta}) = [\mathbf{I} - \mathbf{V}\,\mathbf{B}^{\mathrm{T}}\,(\mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}})^{-1}\,\mathbf{B}]^{\mathrm{T}}\,\mathbf{V}\,(\vec{\theta})\,[\mathbf{I} - \mathbf{V}\,\mathbf{B}^{\mathrm{T}}\,(\mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}})^{-1}\,\mathbf{B}]\ .$$

If the projection operator in brackets is symmetric and therefore orthogonal, this reduces to

$$= \mathbf{V} - \mathbf{V}\mathbf{B}^{\mathrm{T}}\,(\mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}})^{-1}\,\mathbf{B}\mathbf{V}\ .$$

These errors are represented by the projection of the error ellipse upon the constraint surface. Because of the condition represented by the constraint, the new errors are less than or equal to the old, i.e., the diagonal terms $V_{ii}(\hat{\theta}) \leqslant V_{ii}(\vec{\theta})$. The off-diagonal terms, the correlations, may be increased or decreased. In particular, if the measurements θ_i were uncorrelated, the $\hat{\theta}_i$ may now be correlated. On the figure above we see how the k = 1 ellipse, corresponding to $\pm\sigma_i$ for the multi-variate Gaussian representing the original measurements, is projected onto the constraint plane.

The projection operation, which is linear, preserves the probability. That is, in two dimensions as in the figure, the projection operator **A** projects a whole line of points in (θ_1, θ_2) space onto a single point on the constraint surface (on the figure, a line). If we integrate over the probability ellipses of the original measurements and obtain the probability for $\vec{\theta}_{\text{true}}$ to lie between any two such lines, that total probability is preserved on the constraint surface. On that surface, it still represents the probability to find $\vec{\theta}_{\text{true}}$ between the two boundaries represented by the limits of integration. The usual one-standard deviation limits are the projections of the extremes of the k = 1 ellipse, and so on. If the constraint surface is perpendicular to one of the axes, the constrained variance in that direction is zero.

This type of problem is encountered when we have a set of measurements which we want to improve by imposing the constraints implied by a certain model. For example, we have a set of momentum measurements and we wish to try as a hypothesis a certain reaction. Conservation of momentum provides constraints which must be satisfied. If, after using some of the conservation equations to calculate any missing quantities, we have one or more of the conservation equations left over, the problem is sometimes referred to as "over-determined". Then imposing the constraints enables us to reduce our errors by the above procedure. It also guarantees that the solution will satisfy the model.

The Non-Linear Problem

Let us consider a more general case. Let us suppose we have a set of momentum and angle measurements of the tracks on an event, y_i, i = 1,...,n with full variance matrix **V**. In addition, there are some ℓ unknown momenta and angles, η_i, and a set of k constraint equations expressing momentum *and energy* conservation at each vertex (so now the problem will be non-linear). We will use a general notation not specifically tied to the kinematics example; the results will be very general. We will write the constraint equations in the form

$$f_j(y_1, \ldots, y_n, \eta_1, \ldots, \eta_\ell) = 0 \, , j = 1, \ldots, k; \; k \geqslant \ell$$

(not necessarily linear in the y's and η's).

Note that, in our kinematic problem, expressing conservation of energy requires making an assumption about the masses. In fact, in practical cases, the momentum measurement often depends to some extent on the particle mass, through the rate of energy loss.

We want to find the values $\hat{y}$ and $\hat{\eta}$ such that

$\chi^2 = (y - \hat{y})^T \mathbf{V}^{-1} (y - \hat{y}) = \text{minimum}$ *and* $f(\hat{y}, \hat{\eta}) = 0$. Note that the η's do not contribute directly to the χ^2. We could opt to find all ℓ of the η's in terms of some of the y's using the constraint equations, assuming that the constraint equations will do that (they may not, in which case, we can proceed no further). Then we could find $k - \ell$ of the y's in terms of the other y's using the rest of the constraint equations; then χ^2 would come from the differences between these "solved" y's and the measurements of them. But this leaves some of the y's at their measured values and relatively large changes in the others; since χ^2 depends upon the squared deviations, we can do better (i.e., lower χ^2) by spreading out the changes among all the y's. The method of Lagrange's multipliers treats all parameters on an equal footing. Introduce k new unknowns λ_i, i = 1, ... k, one for each constraint. Our unknowns are now

$n\ y_i$'s with n×n variance matrix V
$\ell\ \eta_i$'s
$k\ \lambda_i$'s corresponding to k f's.

We minimize $\chi^2 = (y - \hat{y})^T V^{-1} (y - \hat{y}) + 2\, \lambda^T f(\hat{y},\hat{\eta})$:

$$\frac{\partial \chi^2}{\partial y} = -2\, V^{-1} (y - \hat{y}) + 2\, \lambda^T \frac{\partial f}{\partial y} = 0$$

$$\frac{\partial \chi^2}{\partial \eta} = 2\, \lambda^T \frac{\partial f}{\partial \eta} = 0 = 2 \left(\frac{\partial f}{\partial \eta} \right)^T \lambda \quad \text{(the last form gets rid of the inconvenient transpose of the lambda)}$$

$$\frac{\partial \chi^2}{\partial \lambda} = 2\, f(\hat{y}, \hat{\eta}) = 0 \,.$$

Remember that $\partial\chi^2/\partial y$ is an n-component vector, etc.

Since the f's are not necessarily linear, we solve by iteration. The key step is to "linearize" f, by expanding it in a Taylor series about the current values of $\hat{y}$ and $\hat{\eta}$, then solve this linear problem as in the preceding problem for improved values, then repeat until no change is observed in the χ^2 from one step to the next, or some other convergence criterion is satisfied. We must start, of course, by finding values for the unknowns $\hat{\eta}$. We can start at any reasonable values found from a subset of the constraint equations, and hope that the procedure converges. If it doesn't converge very quickly, which can be because the problem is too non-linear and we are far from the constraint hypersurface, we may assume that the ultimate solution will be of high χ^2 (and therefore fail a goodness-of-fit test) and we terminate the iteration. We do this because unsuccessful iterations consume vast amounts of computer time.

In more detail, the iteration procedure can be as follows (techniques which don't rely explicitly on evaluation of the derivatives of the χ^2 are also very popular, although they are often slower than this more direct method—if the problem is well-behaved). We start by finding reasonable starting values for the unknown η_i's, using ℓ of the constraint equations as convenient.

The final result is sometimes sensitive to the starting values for both y and η. They may be so far wrong that the non-linearities of the problem mean that a good fit cannot be found. Another problem is that the χ^2 may be a complicated function of the variables, with more than one minimum. In some cases the starting values are close to a local minimum which is not the lowest minimum actually available; the iteration tends to find the local minimum rather than the global minimum.

Expand f(y, η) in a first-order Taylor series about the current trial values for y and η; let us call them y_c and η_c. Let $\hat{y}$ and $\hat{\eta}$ be new values which are reasonably close to y and η, and which we hope will prove to be points at which the constraints will be satisfied (f = 0) with the lowest χ^2. We attempt to predict $\hat{y}$ and $\hat{\eta}$ from the expansion:

$$f_j(\hat{y},\hat{\eta}) \simeq f_j(y_c,\eta_c) + \sum_{i=1}^{n} \frac{\partial f_j}{\partial y_i}\bigg|_{y_i^c} (\hat{y}_i - y_i^c) + \sum_{i=1}^{\ell} \frac{\partial f_j}{\partial \eta_i}\bigg|_{\eta_i^c} (\hat{\eta}_i - \eta_i^c) \,; \quad j = 1,\ldots, k \,.$$

Setting this equal to 0.0 gives linear constraints on $\hat{y}$ and $\hat{\eta}$. There are only $k - \ell$ independent constraints here, because ℓ of the constraint equations are used to get the η's (i.e., ℓ of the constraint equations are satisfied by construction of the η's). This is a linear approximation to f, which we can use in the manner of the linear constraint equation of the linear problem. That constraint equation in the form $\mathbf{B}\vec{\theta} = b$ is recovered (because of the complication of the ℓ unknowns, we don't recover exactly the same equations) if we write $\vec{\theta}^T = (y_1,..., y_n, \eta_1,..., \eta_\ell)$ (remember the convention that vectors are written in column form), and the j^th row of **B** as $\left(\frac{\partial f_j}{\partial y_1}, \ldots, \frac{\partial f_j}{\partial \eta_\ell}\right)$. **B** has $k - \ell$ rows and $n + \ell$ columns. Then, the constant b is a $k - \ell$ component vector with element $b_j = -f_j(y_c,\eta_c) + \mathbf{B}\vec{\theta}_c)_j$. Let us proceed.

The plan is to find $\hat{y}$ and $\hat{\eta}$, the improved values of y and η, satisfying the linearized constraints above: $f_j(\hat{y},\hat{\eta}) = 0$. We may have to iterate the following procedures several times until the derivative of the χ^2 is within a specified tolerance of 0.0 and simultaneously the value of $f(\hat{y},\hat{\eta})$ (from an exact calculation, not the Taylor series prediction) is within a certain tolerance of 0.0. To carry out this plan we write, using the rest of the normal equations,

$$\mathbf{V}^{-1}(y - \hat{y}) = \left(\frac{\partial f}{\partial y}\right)^T \lambda \text{ , and therefore } \hat{y} = y - \mathbf{V}\left(\frac{\partial f}{\partial y}\right)^T \lambda$$

which will yield the improved value for $\hat{y}$ once we have the vector λ. The y that appears here is still the original measured y. The derivatives are evaluated at the current y_c and η_c

At this point we recognize that in the kinematics problem and perhaps many others, the variance matrix **V** may not be a constant throughout the space containing y and η, but in general depends upon the true values of y and η. At each stage of the iteration, therefore, we should in this case replace **V** by our current best estimate of **V**, using y_c and η_c. If **V** varies rapidly over the space between the measured y and the final $\hat{y}$ satisfying the constraints, our stepping procedure may fail to improve the χ^2; indeed it may get worse. However, it is always true that the χ^2 will improve over a small enough distance $y_c - \hat{y}$, over which distance **V** (and f as well) is close enough to a constant. Therefore, we may decide to reduce the size of the step below that predicted by the procedure we are in the course of demonstrating, if we find the χ^2 has worsened. Many algorithms exist to attempt to find the swiftest convergence in highly non-linear cases like this.

Let us complete the problem.

$$\hat{y} = y - \mathbf{V}\left(\frac{\partial f}{\partial y}\right)^T \lambda \text{ ,}$$

which means that

$$f(\hat{y},\hat{\eta}) \simeq 0 \cong f(y_c,\eta_c) + \mathbf{F}_y(y - \mathbf{V}\mathbf{F}_y^T\lambda - y_c) + \mathbf{F}_\eta(\hat{\eta} - \eta_c)$$

where **F** is the matrix of derivatives of f, evaluated at the current best values of y and η:

$$(\mathbf{F}_y)_{ij} = \left.\frac{\partial f_i}{\partial y_j}\right|_{y_c}, \ (k \times n); \ (\mathbf{F}_\eta)_{ij} = \left.\frac{\partial f_i}{\partial \eta_j}\right|_{\eta_c}, \ (k \times \ell)$$

In these expressions, y_c is the current best guess at the true value of y; at the outset this is identical with y but on subsequent steps is replaced by the value of y obtained on the previous step, namely, the $\hat{y}$ we found the last iteration.

We now have a matrix equation involving only λ and $\hat{\eta}$ as unknowns. As this equation represents k coupled linear equations, it is not by itself adequate for the k λ's plus the ℓ $\hat{\eta}$'s. We can, however, now express the λ's in terms of known quantities and the $\hat{\eta}$'s:

$$f(y_c,\eta_c) + \mathbf{F}_y(y - y_c) + \mathbf{F}_\eta(\hat{\eta} - \eta_c) \simeq \mathbf{F}_y\mathbf{V}\mathbf{F}_y^T\lambda; \text{ therefore}$$

$$\lambda \simeq (\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}[f(y_c,\eta_c) + \mathbf{F}_y(y - y_c) + \mathbf{F}_\eta(\hat{\eta} - \eta_c)] ,$$

where the k × k symmetric matrix $\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T$ is assumed to be non-singular. We note that λ is no longer a constant but varies with the fit, a consequence of the non-linearity. When we began, we used ℓ of these equations to find starting values for the η's. But at this stage we do something different. We can solve for the $\hat{\eta}$'s using the ℓ equations

$$\frac{\partial\chi^2}{\partial\eta} = 2\mathbf{F}_\eta^T\lambda = 0 \text{ ; inserting } \lambda \text{ above we get}$$

$$\hat{\eta} = \eta_c - [\mathbf{F}_\eta^T(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}\mathbf{F}_\eta]^{-1}\{\mathbf{F}_\eta^T(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}[f(y_c,\eta_c) + \mathbf{F}_y(y - y_c)]\} ,$$

and finally $\hat{y} = y - \mathbf{V}\mathbf{F}_y^T\lambda$.

Note that, upon convergence, $f(y_c, \eta_c) \simeq 0$, in the above expression. We then test for convergence of χ^2 to a minimum within a pre-selected tolerance. We could instead check and see if y and η have changed by more than a specified amount (or a specified %, but in this case we must be careful in the event one or more of the y's or η's is near zero), and use that as a test for convergence.

Only after convergence can we make a test for goodness-of-fit (we could decide to make preliminary tests before convergence, to see if we are making satisfactory progress). This test, possible only if the y's have a known p.d.f., which we will here assume to be normal, consists of comparing the final χ^2 with the p.d.f. for a χ^2 of $k - \ell$ degrees of freedom. If the probability of getting a worse χ^2 than the one we observed is too small, we might assume that something is wrong with the data, model, or both. The meaning of "too small" is not well defined. In the kinematic case, one often rejects if $P(\chi^2 > \chi^2_{obs}) < 1\%$, but that is because other hypotheses may be readily available, and also, it is known that the various momenta are easily mismeasured. However, before accepting or rejecting a model, it is worthwhile to check the residuals, $(y_i - \hat{y}_i)/\sqrt{V(y_i) - V(\hat{y}_i)}$ (also called "pulls" or "stretches" in cases like this where the measured quantities are the ones of direct physical interest). As before, if the number of degrees of freedom (also called, paradoxically, the "number of constraints", since it represents the number of constraint equations in excess of the number needed to determine the unknown η's, i.e., the degree to which the problem is overdetermined) $k - \ell$ is equal to 1, the condition for $\chi^2 + 2\,f\,\lambda$ to be minimum implies that all the residuals are equal in absolute value (provided the correlations among the y_i's are negligible).

If the number of constraints is 0 ($k = \ell$), then we can just determine the unknown η's, and the "best fit" values of y_i consist of the measured values — no improvement is possible from this source. The χ^2, of course, is zero. However, there may exist $k = \ell$ type cases in which the measured y's and calculated η's are outside the boundary of a physical region. In this case, one might decide to improve the measurements by moving them to the nearest (lowest χ^2) point on the boundary of the physical region. Again, a χ^2 test at the conclusion of this procedure should be made to see if this is at all reasonable.

The ability of the χ^2 test to discriminate among models should not be taken too seriously. If the variance matrix **V** is constant throughout the region of interest, and if the measurements are all perfectly Gaussian, one can have faith in the probabilities the χ^2's imply. This is rarely more than approximately true in practice, and one should be aware of the possibility that, in individual cases, this faith may be grossly misplaced.

Regarding errors, our best guess for the dependence of $\hat{y}$ and $\hat{\eta}$ upon y comes from the last step in the iteration. This usually involves small changes in y and η from which we can estimate reasonable derivatives. We regard the next-to-last values of y and η as being merely parameters in the dependence of $\hat{y}$ and $\hat{\eta}$ upon y. Then we approximate $\mathbf{F}_y$, $\mathbf{F}_\eta$, and $\mathbf{V}$ as constants, evaluated at $\hat{y}$ and $\hat{\eta}$, and use propagation of errors. The validity of the errors calculated this way depends upon the validity of these approximations. Write

$$\hat{y} = g(y) \,;\, \hat{\eta} = h(y) \,;$$

g and h being functions which we know, as follows. First,

$$g = y - \mathbf{V}\mathbf{F}_y^T\lambda \quad [g \text{ is } (n\times 1)] \quad [\text{Note: } \mathbf{V}(y) \equiv \mathbf{V}] \,.$$

$$= y - \mathbf{V}\mathbf{F}_y^T\,(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}\,[f(y_c,\eta_c) + \mathbf{F}_y(y - y_c) + \mathbf{F}_\eta(\hat{\eta} - \eta_c)] \,.$$

The values of y_c and η_c are the known former $\hat{y}$ and $\hat{\eta}$ of the next-to-the-last step. We must substitute now for $\hat{\eta}$:

$$g = y - \mathbf{V}\mathbf{F}_y^T(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}\,\{\mathbf{I} - \mathbf{F}[\mathbf{F}_\eta^T\,(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}\mathbf{F}_\eta]^{-1}\mathbf{F}_\eta^T(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}\}$$

$$\times\,[f(y_c,\eta_c) + \mathbf{F}_y(y - y_c)]$$

Similarly,

$$h = \eta_c - [\mathbf{F}_\eta^T(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}\mathbf{F}_\eta]^{-1}\,\{\mathbf{F}_\eta^T\,(\mathbf{F}_y\mathbf{V}\mathbf{F}_y^T)^{-1}\,[f(y_c,\eta_c) + \mathbf{F}_y(y - y_c)]\} \,;$$

h is $(\ell \times 1)$. Then

$$\mathbf{V}(\hat{y}) = (\frac{dg}{dy})^T\,\mathbf{V}(y)(\frac{dg}{dy}) \,;\, (n \times n)$$

$$\mathbf{V}(\hat{\eta}) = (\frac{dh}{dy})^T\,\mathbf{V}(y)\,(\frac{dh}{dy}) \,;\, (\ell \times \ell)$$

$$\mathrm{COV}(\hat{y},\hat{\eta}) = (\frac{dg}{dy})^T\,\mathbf{V}(y)\,(\frac{dh}{dy}) \,;\, (n \times \ell) \,.$$

The derivative matrices $\frac{dg}{dy}$ and $\frac{dh}{dy}$ are $(n \times n)$ and $(\ell \times n)$, respectively. Write

$$\mathbf{P} = \mathbf{F}_y^T\,(\mathbf{F}_y\,\mathbf{V}\mathbf{F}_y^T)^{-1}\,\mathbf{F}_y \;;\; (n \times n)$$

$$\mathbf{Q} = \mathbf{F}_y^T\,(\mathbf{F}_y\,\mathbf{V}\mathbf{F}_y^T)^{-1}\,\mathbf{F}_\eta \;;\; (n \times \ell)$$

$$\mathbf{S}^{-1} = \mathbf{F}_\eta^T\,(\mathbf{F}_y\,\mathbf{V}\mathbf{F}_y^T)^{-1}\,\mathbf{F}_\eta \;;\; (\ell \times \ell) \,.$$

Then

$$\mathbf{V}(\hat{y}) = \mathbf{V}(y)\,[\mathbf{I} - (\mathbf{P} - \mathbf{Q}\mathbf{S}\mathbf{Q}^T)\,\mathbf{V}(y)] \,;$$

$$\mathbf{V}(\hat{\eta}) = \mathbf{S} \,;\; \text{and}$$

$$\mathrm{COV}(\hat{y},\hat{\eta}) = -\mathbf{V}(y)\mathbf{QS}\,.$$

The residuals $\vec{r} = y - \hat{y}$ have variance matrix

$$\mathbf{V}(\vec{r}) = \mathbf{V}(y) + \mathbf{V}(\hat{y}) - 2\,\mathrm{COV}\,(y,\hat{y})$$

$$\simeq \mathbf{V}(y) - \mathbf{V}(\hat{y}) = \mathbf{V}(y)(\mathbf{P} - \mathbf{QSQ}^{\mathrm{T}})\,\mathbf{V}(y)\,.$$

We could also have obtained these relations from

$$V_{ij}(y) = 2\left(\frac{\partial^2\chi^2}{\partial y_i \partial y_j}\right)^{-1}\,, \text{ etc.}$$

If the linearity assumptions are strongly violated, the χ^2 will have significant higher derivatives, and we can use the graphical method described before: construct the plane $\chi^2 = \chi^2_{\mathrm{min}} + a$, where $a = 1, 4, 9$, etc., and determine the extreme values of each of the y's and η's which intersect this plane. This is very time-consuming, and in the case in which large numbers of such fits must be done, as in the kinematics example, one usually doesn't do this.

Finally, if the problem is strongly non-linear, this technique for finding the χ^2 minimum via successive approximations to the derivatives may "blow up" (i.e., find a "local" minimum at infinity rather than the global minimum nearby), take an unreasonable amount of time to converge, find an unphysical solution at some local minimum in a disallowed region (e.g., negative momenta), or suffer from some other problem. In this case a slow-but-sure minimization is recommended which calculates χ^2 at selected points about the current trial values of the parameters and proceeds carefully using that information.

Selected Bibliography

General Books with a Practical Orientation

Bevington, Philip R., *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, London, 1969).

Devore, Jay L., *Probability and Statistics for Engineering and the Sciences* (Brooks/Cole, Monterey, California, 1982).

Eadie, Drijard, James, Roos, and Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, London, 1971).

Frodesen, Skjeggestad, and Tøfte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Oslo, 1979).

James, F., "Probability, Statistics, and Associated Computing Techniques," in *Techniques and Concepts of High Energy Physics II*, Thomas Ferbel, ed., (Plenum Press, New York, 1983).

Meyer, Stuart L., *Data Analysis for Scientists and Engineers* (J. Wiley and Sons, London, 1975).

Pollard, J.H., *A Handbook of Numerical and Statistical Techniques* (Cambridge, London, 1979).

The Same, Emphasis on Multivariate Analysis

Green, Paul E., *Analyzing Multivariate Data* (Dryden Press, Hinsdale, Illinois, 1978).

General Books with a Theoretical Orientation — not for the timid —

Bickel, Peter J. and Doksum, Kjell A., *Mathematical Statistics, Basic Ideas and Selected Topics* (Holden-Day, San Francisco, 1977)

Cox, D.R., and Hinkley, D.V., *Theoretical Statistics* (Chapman and Hall, London, 1979).

Hogg, R.V. and Craig, A.T., *Introduction to Mathematical Statistics* (Macmillan, London, 1965 or later editions in paperback)

Lehmann, E.L., *Theory of Point Estimation* (John Wiley and Sons, New York, 1983).

Lindgren, B.W., *Statistical Theory* (Collier Macmillan, London, 1976).

Specializing in Least-Squares Techniques

Draper, Norman and Smith, Harry, *Applied Regression Analysis, Second Edition* (John Wiley and Sons, New York, 1981).

Neter, John and Wasserman, William, *Applied Linear Statistical Models* (Richard D. Irwin, Inc., Honewood, Illinois, 1974).

Rao, C. Radhakrishna, *Linear Statistical Inference and its Applications* (John Wiley and Sons, New York, 1973).

Specializing in Monte Carlo Techniques

Rubinstein, R.Y., *Simulation and the Monte Carlo Method* (John Wiley and Sons, New York, 1981).

Specializing in Robust Estimation

Conover, W.J., *Practical Nonparametric Statistics* (John Wiley and Sons, New York, 1980).

Rey, William J.J., *Introduction to Robust and Quasi-Robust Statistical Methods* (Springer-Verlag, New York, 1983).

Specializing in Outlier Identification

Hawkins, D.M., *Identification of Outliers* (Chapman and Hall, New York, 1980)

Useful Mathematical References

Abramowitz, Milton and Stegun, Irene A., *Handbook of Mathematical Functions* (Dover, New York, 1972).

Bartsch, Hans-Jochen, *Handbook of Mathematical Formulas* (Academic Press, New York, 1974).

Basilevsky, Alexander, *Applied Matrix Algebra in the Statistical Sciences* (North-Holland, New York, 1983).

Burington, Richard Stevens, *Handbook of Mathematical Tables and Formulas* (Handbook Publishers, Inc., Ohio, 1958 or later editions).

And, of Course

Kendall, M.G. and Stuart, A., *Advanced Theory of Statistics* (Hafner, New York, Vol. I, 4th ed. 1977, Vol. II, 4th ed. 1979, Vol. III, 3rd ed. 1968)

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720