Phil 29

Lecture Notes
Statistics for Physicists - Eadie, Drijard, Roos

Advanced Theory of Statistics - Kendall & Stuart

Handbook of the Common Distributions, Haight

Selected Papers on Noise and Stochastic Processes - various authors

included famous S.O. Rice paper on noise

several RMP papers on statistical physics

Ashley - On the theory of stochastic processes and their application to the theory of cosmic radiation


On probability problems in the theory of cosmic rays

Stochastic Processes - Parzen
Decision Theory  
Student's F test  
Goodness of Fit  
Distribution independent tests

Generation of Monte (Pseudo) Random Numbers  
Monte Carlo Integration

Things I don't understand very well:  
Minimal Spanning Tree: Investigates Clustering

Topics requested Sam

Very little on how to win at Las Vegas e.g. combinatoric analysis.

Course  
Either Homework (> 50%) + Oral  
or Prepare discussion of some advanced topic at end of course (either I or student will present).
Books:

For the first part of course (time
decision theory), choose between

Statistical and Computational Methods
in Data Analysis by S. Brandt - North-
Holland.

Statistical Methods in Experimental
Physics by W.T. Eadie et al., North-Holland.

Brandt's book has more examples
(programs) and is simpler. (Beverlin
[old notes] is also good at a low level).

And for the mathematical background, see

Mathematical Methods of Statistics
by H. Cramer (Princeton University
Press).
Subjests:

Laws of Statistics
Basic Properties following from laws:
Bayes Theorem
Continuous/Discrete Random Variables
Characteristic Functions
Sums of Random Variables

Probability Distributions
Central Limit Theorem

Probability Distributions
Biomial
Poisson
Gaussian

Estimation of Parameters
Maximum Likelihood
$\chi^2$
Method of Moments
Bias/Error
Bayes v. Non-Bayes
Robust (Distribution-free methods)
Minimization of Functions.
Physics Logistics

Teacher J. Fox
Room 207 Booth

Please call x3765, x6673 if want to ask questions.
Typically around 8.15am - 5.30pm

Grades:
1) J. Miller
   x 2918
   Room 108 Sloan Annex

2) M. Bucher
   x 4861
   60 W. Bridge

They will alternate problem sets
D. Statistics

D1: Introduction

There are four main topics in statistics or probability theory.

1) "Combinatoric Brilliance": i.e., how to win at Las Vegas.
2) "Design of Experiments": Histograms, variance, skewness etc., e.g., how to take reliable public opinion polls.
3) "Formal Theory": Hausdorff measure, central limit theorem etc.
4) "Parameter Estimation": $\chi^2$, maximum likelihood, method of moments.

Further there are some topics which are really techniques in numerical analysis.

5) Find minimum/maximum of a function
of several variables.

6) Generation of random numbers and pseudo random variables: use of these in Monte Carlo evaluation of integrals.

Most mathematical books cover i) to 3). Typical is

Feller: An Introduction to Probability Theory and Its Applications, Vols 1, 2.

This only covers 1-3 and is very good on 2.

Hoel: Introduction to Mathematical Statistics: is quite readable and covers i) to 4).

Beverton: Data Reduction and Error Analysis for the Physical Sciences: This covers 2, 4 and 5. It is perhaps the best simple handbook of the
formulas. However it is naive on 4), 5) and if you start doing complicated things on $x^2$/maximum likelihood/minimization, Bevington will lead you astray. Another and even shorter general reference is Mathews and Walker: chapter 14: this mentions all the important points.

For 4), there are several notes by physicists:


5. Oscar, ucll-8417 (1958). "Notes on Statistics for Physicists"

5. Yellin, DESY 72/12: Preprint (1972)

K. McDonald, CTSL Internal Report #57

Arnold and MacGregor, Methods in Computational Physics - Vol 6

Annis, Cheston and Primakoff, Rev. Mod. Phys 25, 818 (59)

while very good on (3) and formal aspects of (4) is Cramer: Mathematical Methods of Statistics.
There is also a recent look on more practical aspects of 4) i.e. a heyday of Cramer into physics applications: W. T. Eade, D. Drijard, F. E. James, M. Roos, B. Sadoulet. Statistical methods in Experimental Physics.

6) I will postpone till we do numerical analysis while there is no good reference on 5) — even though if one enquired (and was read) it would save millions of dollars of computer time. Tolerable is:

**Fox:** (not me): Optimization methods for Engineering Design.

**Bad** (i.e. irrelevant mathematics) is

So the plan of course is:

D2: Formal Points: definitions, Bayes Theorem, Gaussians, Central Limit Theorem.

D3: Binomial and Poisson Distributions:

Las Vegas ... > Biology (Birth and Death Processes).

D4: Estimation of Parameters

Maximum likelihood, $\chi^2$, and method of moments, Goodness of fit, minimization techniques.
STATISTICS FOR PHYSICISTS

Geoffrey Fox

I: BASIC TECHNIQUES AND THEORY

- Formal Theory
- Special distributions including those needed for combinatorics
- Statistical Inference
  - Parameter Estimation
  - Decision Theory
D2  Basic Definitions
    Bayes
    Continuous Distributions
        Random Variables
        Joint Distributions
            Means, Moments, Correlations

Central Limit Theorem
Monte Carlo Integration
Extensions to Correlated Functions

D3  Binomial and Poisson Distributions
    Compound Poisson Distribution
    Additivity Property and Gaussian Limits of Binomial and Poisson Distributions
    Birth and Death Processes
D4 Estimation of Parameters

Maximum Likelihood Method

Method and Heuristic Justification

Example

Proof

$\chi^2$ Method

Relation to Maximum Likelihood

Example

Counting (Binned) Experiments

Event Experimental

Method of Moments

D5 Minimization

Basic Derivative Method

Error Calculation

Real World

Eigenvalue and Marquardt
Goodness of Fit

$X^2$

Confidence Level

Barlett's Sudden

Hypothesis Testing

Student's t distribution

Robust Estimation

Generation of Random Numbers

in $[0, 1]$ Uniform

General Distribution
Useful Books for Statistics Part of Course

J.P. Yost  "Lectures on Probability and Statistics" - copies handed out in class (contact G. Fox if you need further information)

W.T. Eadie et al. "Statistical Methods in Experimental Physics" - a little sophisticated for this course but technically excellent

   Chapter 7, 13, 14
   - Computer-related issues
   This has FORTRAN and PASCAL programs available

J. Mathews and R.L. Walker (Benjamin) "Mathematical Methods of Physics"
   Chapter 14
D2 Formal Points

D2/01 Formal Definitions

(i) A major difficulty in formulating the theory of probability is that one must relate it to the real world. Cramér (Mathematical Methods of Statistics - Princeton) amusingly notes that some of the axiomatic approaches to probability are comparable to definitions of geometry in terms of the fundamental idea of a point as a limit of a spot of chalk dust of smaller and smaller size.

I will first dispose of my 24 lectures on formal probability by noting that the theory can be based on the following axioms.

Consider a variable point \( x \) in a space \( Y \) where \( Y \) is some subset of \( \mathbb{R}^k \).
Further consider the family of all Borel sets $S$ in $Y$. You will of course recall that a Borel set is gotten by performing (finite or denumerably many times) on intervals $[a, b], (a, b), (a, b], [a, b], \ldots$ the operations of addition, subtraction, or multiplication. (and continuing process denumerably often).

\[ \text{modern math: } \Rightarrow \]

\[ S_1 - S_2 S_2 \]

\[ S_1 + S_2 = \text{whole thing.} \]

**Axiom 1:** To every $S$ corresponds a non-negative number $P(S)$ which is called the probability of the event $X$ belonging to $S$.

**Axiom 2:** $P(Y) = 1$

**Axiom 3:** $P(S)$ is a completely additive
Set function:

\[ P(S_1 + S_2 + \ldots) = P(S_1) + P(S_2) + \ldots \]

if the \( S_i \) are disjoint \( \implies S_i \cap S_j = \emptyset \).

Note that it follows from the axioms that \( c) 0 \leq P(S) \leq 1 \) for any \( S \). \( \text{Proof: } P(S) > 0 \) was posulated. Take \( S_1 = S, S_2 = Y - S = S^* \) (the complement of \( S \)) in axiom 3.

\[ P(S_1) + P(S_2) = P(Y) = 1 \]

but \( P(S_2) > 0 \implies P(S_2) = P(S) \leq 1. \]

(b) \( P(S_1 + S_2) \leq P(S_1) + P(S_2) \), any \( S_1, S_2 \) with equality if \( S_1 \) and \( S_2 \) are disjoint.

\[ \text{Proof: } S_1 + S_2 = (S_1 - S_2) + S_2 \]

\[ P(S_1 + S_2) = P(S_1 - S_2) + P(S_2) \text{ as sets disjoint. But } P(S_1 - S_2) \leq P(S_2) \]

-See (a). This proves desired result.

Anyhow the variable point \( X \) is now called a random variable and \( P(S) \) is the probability function of \( X \).
To relate this definition to the outside world, we must resort to the vanishing chalk dust approach as espoused by Mathews and Walker.

If we flip an (unbiased) coin \( N \) times, we expect

\[
\lim_{N \to \infty} \frac{N_h}{N} = \frac{1}{2}
\]

where \( N_h \) is the number of times it turns up heads. We define:

\[
P(\text{heads}) = \lim_{N \to \infty} \frac{N_h}{N} = \frac{1}{2}
\]

This is a random variable with only two possible values - heads or tails. Alternatively, we could say \( x = 0 \) and 1. Then in our present language any Borel set containing 0 and not 1 has \( P(S) = \frac{1}{2} \), any containing 0 and 1, \( P(S) = 1 \) etc.
**Bayes Theorem: Conditional Probabilities**

Take two possible (Borel) subsets of \( X \in \mathbb{R}^p \). Call them A and B.

A could be good \( x = 0 \)

B \( x = 1 \) only.

\[
A - \overline{A}B \quad \overline{A}B \quad B - \overline{A}B
\]

\((A + B)^+\)

Generally, \( Y \) is divided into four subspaces by A and B and following chalk dust method - imagine \( N \) experiments, give the following frequencies:

\[
N_1 = n(A - \overline{A}B)
\]

\[
N_2 = n(\overline{B} - \overline{A}B)
\]

\[
N_3 = n(\overline{A}B)
\]

\[
N_4 = n((A + B)^+) \text{ - complement of } (A + B)
\]

of events \( n(A) \) falling in subspace \( Y \), then \( N_1 + N_2 + N_3 + N_4 = N \).

We previously define \( P(A) \) and now we expect
\[ P(A) = \lim_{N \to \infty} \frac{N_1 + N_3}{N} \]
\[ P(B) = \lim_{N \to \infty} \frac{N_2 + N_3}{N} \]
\[ P(A \cup B) = \lim_{N \to \infty} \frac{N_1 + N_2 + N_3}{N} \]  \hspace{1cm} (D.2.1)

Now we define conditional probability
\[ P(S/A) \] \hspace{1cm} \{ \text{read as probability of } S \text{ given } A \text{ has occurred} \} \hspace{1cm} \text{as} \hspace{0.5cm} P(SA)/P(A). \hspace{0.5cm} \text{Note}\hspace{0.5cm} \text{this is a probability according to}\hspace{0.5cm} \text{axioms if we replace } Y \text{ by } N. \text{\hspace{1cm}} \\
\hspace{1cm} \text{Then in terms of frequencies,} \\
\hspace{1cm} P(A/B) = \frac{N_3}{N_2 + N_3} \hspace{1cm} \text{as } N \to \infty \hspace{1cm} (D.2.2) \]
\hspace{1cm} P(B/A) = \frac{N_3}{N_1 + N_3} \hspace{1cm} \text{as } N \to \infty \\
\hspace{1cm} \text{whence:} \\
\hspace{1cm} P(B) \cdot P(A/B) = P(A) \cdot P(B/A) \hspace{1cm} (D.2.3) \\
\hspace{1cm} \text{which is Bayes' law. Of course} \\
\hspace{1cm} \text{(D.2.3) is obvious from def }; \hspace{0.5cm} \text{we didn't need frequencies to prove} \hspace{0.5cm} \text{it.}
\( n_{\text{tot}} \) pictures with some property of interest
\( n(A) \) Number found by an undergraduate during finals week
\( n(B) \) Number found by a graduate student working on a theory in 26 dimensions

Example (page 11, Eadie et al.)

A: Scan Number One
B: Scan Number Two

We can tag members of \( C = AB \) (A and B).

\[
p(c) = p(A) \cdot p(B) \]

(In practice not true as say short tracks systematically missed!)

\[
\frac{n(c)}{n_{\text{tot}}} = \frac{n(A)}{n_{\text{tot}}} \cdot \frac{n(B)}{n_{\text{tot}}}
\]

Or \( n_{\text{tot}} = \frac{n(A) \cdot n(B)}{n(c)} \) is desired estimate of total number of events.

In practice, second scan will only be on a subset of data sample.

Note if A, B independent

\[
p(A/B) = \frac{p(AB)}{p(B)} = p(A) \quad \text{independent of } B.
\]
Example: page 18 of Yost

\[ B \begin{cases} B_1 = 0 & \text{ill} \\ B_2 = 1 & \text{well} \end{cases} \]

\[ A = \text{"Test Positive"} \]

\[ P(B|A) = \frac{P(B)P(A|B)}{P(A)} \]

Apply separately \( B = B_1 \) or \( B_2 \).

\[ P(B|A) = \frac{\frac{P(B_2)P(A|B_2)}{P(A_1) = .004 \times 99 + .02 \times .999}}{\frac{0.999 \times .02 \delta(x) + .004 \times .99 \delta(1-x)}{\delta(x)}} \]

\[ P(B_1|A) = \text{"ill"} = .047 \]

\[ P(B_2|A) = \text{"well"} = .953 \]
The Bayes Controversy

Now we have derived Bayes Theorem where A and B are (Borel) sets in a space populated by a random variable. Now a useful "application" of the theorem is to take one of them A as the Borel set given a set of experimental results - this is correct. B is a set of hypotheses e.g. in a set in space of values of a theoretical parameter e.g. we measure scattering angles \( \Theta \) and wish to find \( \alpha \) when probability of observing \( \Theta \) is \( \frac{1}{2}(1 + \alpha \cos \Theta) \).

A is a set in \( \Theta \) space.
B is a set in \( \alpha \) space.

Then experimentalists measure \( P(A/B) \) i.e. values of \( \Theta \) subject to known (by the almighty and those on fourth floor) single value of \( \alpha \). Bayes says remarkably

\[
P(B/A) = \frac{P(A/B) P(B)}{P(A)}
\]

and so this gives you what you really want - the "probability distribution of \( \alpha \)" consequent from measurements of \( \Theta \). Problem with this is that:

a) \( \alpha \) is not a random variable - it only takes one value. Answer: Our knowledge of \( \alpha \) can be considered as a random variable.
b) What is $p(B)$? We will discuss this so called a priori knowledge of $B$ later on. It essentially summarizes all previous (other) knowledge of $B$ gained from previous (other) experiments.

c) What is $p(A)$? The probability of measurement of $B$ given any possibility of $A$. Nobody knows what this is! You essentially determine it by normalizing $p(B/A)$ or equivalently note that

$$
p(B_1/A) = \frac{p(A/B_1)p(B_1)}{p(B_2/A) + p(A/B_2)p(B_2)}
$$

are indeed independent of the noxious $p(A)$. 


Continuous Distributions

To be formal and reveal relics of a primordial mathematical training, take random variables in $\mathbb{R}$ and consider special Borel set $\mathcal{E}$ such that it has probability $F(x)$ with $F(\infty) = 1$.

Now in some cases we can differentiate $F(x)$ and so write:

$$F(x) = \int_{-\infty}^{x} p(x) \, dx$$

where $p(x)$ is probability density. One can clearly interpret $p(x) \, dx$ as the probability that random variable $X$ takes values in range $x \leq x + dx$.

(a) In the case of a tossed coin, let heads be $x = 1$ and tails $x = 0$. Then for (an unbiased) coin

$$p(x) = \frac{1}{2} \left[ \delta(x) + \delta(x-1) \right]$$
(a) \( p(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \) (D2.4)

This is the probability density of the normal or Gaussian distribution - so called because it was first discussed by de Moivre in 1733.

\[
\int_{-\infty}^{+\infty} p(x) \, dx = 1
\]

Also the associated distribution function is:

\[
F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{x} dy \exp \left[ -\frac{(y - \mu)^2}{2\sigma^2} \right]
\]

which cannot be done analytically except by relating it to the error integral:

\[
\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \exp(-y^2) \, dy
\]
D2/4: Functions of a Random Variable

If $x$ is a random variable, then so is $y = f(x)$. Let $x$ have probability distribution $p(x)$ and $y = g(y)$. Then probability of $x$ lying in $[x, x+dx]$ is $p(x)\,dx$; let $[y, y+dy]$ be associated interval in $y$.

Clearly $p(x)\,dx = g(y)\,dy$

and $dy/\,dx = |f'(x)|$

So $p(x) = g(f(x))|f'(x)|$ \hspace{1cm} (D2.5)

There are some obvious modifications of this if the $x \leftrightarrow y$ relation is not single-valued.
Joint Probability Distributions

We can now consider the important extension of (12.14) to more than one variable. Suppose we have two random variables \( x \) and \( y \) each taking values in \( \mathbb{R} \). Then we can conceive of a random variable \( z \) which is the ordered pair \((x, y)\). Even if we knew the probability distributions of \( x \) and \( y \)

\[
x \rightarrow p(x) \, dx
\]
\[
y \rightarrow q(y) \, dy.
\]

The distribution of \( z \) is not defined. Suppose that \( z \) has density \( r(x, y) \) so that probability that \( z \) lies in volume element \( dx \, dy \) is \( r(x, y) \, dx \, dy \). A prior, all we know is that \( r \) satisfies

\[
\int r(x, y) \, dy = p(x)
\]
\[
\int r(x, y) \, dx = q(y)
\]

These are called marginal distributions (EDTR5 p.18).
(a) Example: Bayes law for densities.

Let event $X$ lie in $[x, x + dx]$ and $Y$ lie in $[y, y + dy]$.

Let $T(X|Y) = t(x|y) dx$ be the probability of $X$ given $Y$. By definition,

$$T(X|Y) = \frac{Pr.(XY)}{Pr.(Y)} = \frac{t(x,y) dx dy}{q(y) dy}$$

or $t(x|y) = \frac{t(x,y)}{q(y)}$ a conditional distribution (EDTJS page. 18).

Similarly let $S(Y|X) = s(y|x) dy$ be the probability of $Y$ given $X$. By definition:

$$s(y|x) = \frac{r(x,y)}{p(x)}$$

or eliminating $r(x,y)$

$$t(x|y) q(y) = s(y|x) p(x) \quad (D2.6)$$

This is a density version of Bayes law.

(2) Example: $y = f(x)$.

When $y$ is a known function of $x$,

clearly $s(y|x) = s(y - f(x))$
Then using above equ. for $S(y|x)$, we can in this case find

$$r(x, y) = S[y - f(x)] p(x).$$

(c) **Independence:**

If the distribution of $z$ satisfies

$$r(x, y) = p(x) q(y) \quad (D2.7)$$

then we say that $x$ and $y$ are statistically independent.

(d) **Function of 2 Variables**

Suppose $w$ rather than $z$ in $\mathbb{R}_2$, we consider $h(x, y)$ in $\mathbb{R}_2$, a nonmeasurable function of $x$ and $y$. Then $w = h(x, y)$ is also a random variable with distribution $d(w)$ where

$$d(w) = \iint dx dy \ S[w - h(x, y)] r(x, y) \quad (D2.8)$$

This may perhaps be guessed very grettily: it is obvious if you just do the $x$-integral in $(D2.8)$ by $\Sigma$ and use same argument in $D2/4$. 

---

D26 Means - moments etc.

(a) Given a one dimensional distribution:

Mean: \( \langle x \rangle = \int_{-\infty}^{+\infty} x \, p(x) \, dx \)

n'th moment: \( \langle x^n \rangle = \int_{-\infty}^{+\infty} x^n \, p(x) \, dx \)

Expectation: \( \langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x) \, p(x) \, dx \) which follows from n'th moment by Taylor expansion of \( f(x) \). This can be considered as mean of the random variable \( z = f(x) \).

n'th central moment:

\( \langle (x - \langle x \rangle)^n \rangle = \int_{-\infty}^{+\infty} (x - \langle x \rangle)^n \, p(x) \, dx \)

Variance: \( V_x = \langle (x - \langle x \rangle)^2 \rangle = 2^{nd} \text{ central moment} \)

Standard deviation \( \sigma_x \): \( V_x = \sigma_x^2 \) where \( \sigma_x \) is used much more than \( V_x \).

Note \( V_x = \langle x^2 \rangle - 2 \langle x \rangle \langle x \rangle + \langle x \rangle^2 \)

\( = \langle x^2 \rangle - \langle x \rangle^2 \) \( \text{(D2.9)} \)

A result which is used continuously.
(b) $n$-dimensional distribution:

The mean now becomes an $n$-dimensional vector

$$<x_R> = \int d^nx \ x_R \ p(x_1, \ldots, x_n)$$

The variance becomes an $n \times n$ moment matrix $M$

$$m_{R \ell} = \int d^nx \ [x_R - <x_R>] [x_\ell - <x_\ell>] \ p(x_1, \ldots, x_n)$$

Note for we can show - analogously to (D2.9) - that

$$m_{R \ell} = <x_R x_\ell> - <x_R><x_\ell> \ (D2.10)$$

Also note that $M$ is symmetric and positive semi definite. [This follows because $y_M x_R y = y_M x_R - x_R y_M y$ - clearly has a positive $\geq 0$ integrand].

Now the diagonal terms are called variances which are $\geq 0$; their
Square roots are standard deviations again. Further,

\[ k \neq l : \sqrt{S_{kl}} = \sqrt{\frac{M_{kl}}{M_{kk} M_{ll}}} \]

is \( \leq 1 \)

and is called correlation coefficient.

Note that if \( S_{kl} = 0 \) - all \( k \neq l \),
then we say that \( x_1, \ldots, x_n \) are
uncorrelated. Further if \( x_1, \ldots, x_n \) are
independent, then (32.10) shows at
once that they are uncorrelated: the
converse does not follow of course.
(except under special circumstances
\( e.g. x_i \) Gaussianly distributed).

One can define higher order tensors
\( M_{klmn} \ldots \) but these are not too important
as in usual applications \( x_1, \ldots, x_n \) are
functions of \( N \gg n \) observations.
Then central limit theorem - too which
we will come - implies that only the mean and MRE are important asymptotically (in N).

(c) **Use of Moment Matrix**:

Often the results of an experiment on \( x_1, \ldots, x_n \) are quoted in terms of means \( \pm \) errors; the errors are standard deviations \( \sqrt{\text{diagonal elements of moment matrix}} \). Now this is throwing away information - the off-diagonal terms in \( M \) - which can be important. Thus suppose some years after the experiment, a theorist decides that \( y = T x \) (\( T \) an \( n \times n \) matrix) is useful. What are errors on \( y \)?

\[
<y_{1}y_{2}> = T_{12} \quad <x_{1}x_{2}> T_{12}^T
\]

or \( M_y = TM_x T^T \) \hspace{1cm} (D2.11)
To find diagonal terms in $E^y$ - i.e. the errors on the axes $y$ - requires off diagonal terms in $M_x$. An example from high energy physics is in the decay of a spin 1 particle, one can define angular distribution by density matrix elements:

$$x_1 = 800$$
$$x_2 = 811$$
$$x_3 = 810$$

in some Lorentz frame. Some years ago these were always given in one frame - the so-called $t$ channel frame. Recently it was found that the values in another frame (the "s channel") were more meaningful physically. Unfortunately certain values $y$ were related precisely such a linear transformation $y = T x$ and one can no longer estimate errors.
The $x_i$ and $y_i$ in this example correspond to expressing angular decay

$$W(\theta, \phi) = 1 + \sum_{l=1}^{3} x_l \, X_l(\theta, \phi)$$

or

$$W(\theta, \phi) = 1 + \sum_{l=1}^{3} y_l \, Y_l(\theta, \phi)$$

where $y_l$ are linear combinations of the $x_i$ corresponding to the $X_l(\theta, \phi)$ being linear combinations of the $Y_l(\theta, \phi)$.

The details are different, but you can see the essential points from

\[
\begin{align*}
X_1 &= \cos \theta & Y_1 &= P_2(\cos \theta) \\
X_2 &= \cos^2 \theta & Y_2 &= P_2(\cos \theta) \\
X_3 &= \cos^3 \theta & Y_3 &= P_3(\cos \theta)
\end{align*}
\]
More examples of correlations

1) Measurement of tracks with chambers

\[ x = a + b \cdot z \]

Track is \[ x = a + b \cdot z \]

slope

intercept

\[ b = \frac{(x_2 - x_1)}{(z_2 - z_1)} \]

\[ a = \frac{z_2 \cdot x_1 - z_1 \cdot x_2}{z_2 - z_1} \]

\( x_i \) are independent random variables but \( a \) and \( b \) are clearly correlated

\[ \langle (b - \langle b \rangle)(a - \langle a \rangle) \rangle = -\frac{z_2 \cdot \sigma_1^2 - z_1 \cdot \sigma_2^2}{(z_2 - z_1)^2} \]

where \( x_i \) has standard deviation \( \sigma_i \)

(Note if \( \sigma_1 = \sigma_2 \) and \( z_1 = -z_2 \) i.e. origin is mid-point between chambers) then \( a \) and \( b \) are uncorrelated. However the origin \( z = 0 \) (in our experiment) was a convenient surveyor's mark and so
was not chosen to minimize correlation! In the case of many chambers the slope and intercept will still be correlated and there will be no simple way to get rid of it. In practice it is important to keep full correlation matrix e.g. best vertex.

2) Measurement of Regge pole intercept

In $E350,$

$$\frac{d\sigma}{dt|dxF} = g(t) (1-x)^{1-2\alpha(t)}$$

where $0 < 1$ ($0.5 \leq x \leq 1$ in experiment)

$g(t)$ is unknown fixed (at fixed $t$) constant of little interest.

$\alpha(t)$ is interesting.

In a (non-linear) fit
$g(t)$ and $\alpha(t)$ are strongly correlated

So that $\int_0^1 g(t) (1-x)^{1-2\alpha(t)}$ is approximately preserved. Unfortunately we only want $\alpha(t)$ (predicted by theory) and cannot use fact that a particular linear combination of $\alpha(t)$ and $g(t)$ is predicted much better than either one.
In practice - when off diagonal items are not given - one assumes they are zero - i.e. variables are uncorrelated. This gives:

$$\sigma_{yi}^2 = \sum_{j=1}^{n} (T_{ij})^2 \sigma_{xj}^2$$

which is the usual rule of addition of errors in quadrature. One has to emphasize that this rule is only valid when $x_j$ uncorrelated.

(d) Error in Lack of Correlation Assumption:

If $y = a_1 x_1 + a_2 x_2$

Then $\sigma_y^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + 2a_1 a_2 \rho \sigma_1 \sigma_2$

and all one can say in general is that:

$$\sigma_y^2 \leq (|a_1 \sigma_1| + |a_2 \sigma_2|)^2$$

The uncorrelated estimate is

$$\tilde{\sigma}_y = \sqrt{a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2}.$$
and the line $\sigma_y$ satisfies

$$0 \leq \sigma_y \leq \sqrt{2} \sigma_y$$

where extremes are only attained when $|a_1,0,1| = |a_2,0,1|$ which is most
dangerous case. Other ratios $|a_1,0,1|/|a_2,0,1| \neq 1$ are better e.g. $a_1,0,1 = 0 \Rightarrow \sigma_y = \sigma_y$!

For $n$ variables, all one can say is

$$0 \leq \sigma_y \leq \sqrt{n} \sigma_y$$

which is clearly pretty bad for

large $n$.

(2) Parameters of a Gaussian Distribution

Suppose $x$ is Gaussian i.e.

$$p(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Then

$$\langle x \rangle = \int_{-\infty}^{+\infty} x p(x) \, dx$$

$$= \mu + \int_{-\infty}^{+\infty} (x-\mu) p(x) \, dx = \mu$$
as last integral clearly vanishes from antisymmetry. So \( \mu \) is mean.

\[
\text{Variance} = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{\infty} (x-\mu)^2 f(x) \, dx
\]

\[
= \sigma^3 \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{\infty} dy \, y^2 \exp \left[ -\frac{y^2}{2\sigma^2} \right]
\]

\[
= \sigma^3 \frac{1}{\sqrt{2\pi} \sigma} \, \frac{\partial}{\partial \sigma} \left\{ \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}} \, dy \right\}
\]

\[
= \sigma^2 \frac{1}{\sqrt{2\pi}} \, \frac{\partial}{\partial \sigma} \left[ \sqrt{2\pi} \sigma \right]
\]

\[
= \sigma^2
\]

So \( \sigma \) is Standard deviation - as its nomenclature suggested.

(f) Addition of Independent Random Variables

Suppose \( y = \frac{1}{n} \sum_{i=1}^{n} x_i \) where the \( x_i \) are independent and all have the same distribution with mean \( \mu \) and standard deviation \( \sigma \). Then clearly \( <y> \) is also \( \mu \).
\[ \langle (y-p)^2 \rangle = \frac{1}{n^2} \sum_{i,j} \langle (x_i-p)(x_j-p) \rangle = \sigma^2 / n \] as all off-diagonal terms vanish.

So exactly, \( y \) has mean \( p \) and std. deviation \( \sigma / \sqrt{n} \). Later we will show that as \( n \to \infty \) (i.e. not exactly) all other moments of \( y \) vanish faster than \( 1/n \) and so \( y \) becomes Gaussian.
Central Limit Theorem

(i) Moment Generating Function

This which is also called characteristic function is defined by

\[ \psi(k) = \int_{-\infty}^{\infty} dx \, e^{ikx} \, p(x) \]

\[ = 1 + ik\langle x\rangle - \frac{k^2}{2!} \langle x^2\rangle + \ldots \]

(a) Note the Fourier inversion theorem gives

\[ p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi(k) \, e^{-ikx} \, dk \]

(b) Most importantly put \( h = x + y \) where \( x \) and \( y \) are independent. From D2/5 it follows that distribution density of \( h \) is

\[ w(h) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(h - x - y) \, p(x) \, q(y) \, dx \, dy \]

So

\[ \psi_h(k) = \int_{-\infty}^{\infty} dh \, w(h) \, e^{ihk} \]
\[ = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{ik(x+y)} p(x) q(y) \, dx \, dy \]

\[ = \phi_x(k) \phi_y(k) \]

So characteristic functions multiply for addition of two random variables. This result can obviously be extended to sum of \( n \) independent random variables.

\[ (c) \quad \phi(k) = \sum_n k^n \frac{E((1x)^n)}{n!} \]

\[ \phi(k) = 1 + l \langle x \rangle k - \frac{\langle x^2 \rangle k^2}{2} \text{ etc.} \]

For a Gaussian distribution

\[ \phi(k) = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{+\infty} \exp(ikx) \exp \left[ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right] \]

\[ = \exp \left\{ i \mu k - \frac{1}{2} k^2 \sigma^2 \right\} \]
(d) Define the cumulant as

\[ K(k) = \ln \phi(k) \]

I have never used this!

Note that

\[ K(k) = \langle x \rangle i k - \left( \langle x^2 \rangle - \langle x \rangle^2 \right) \frac{k^2}{\sigma^2} \]

\( \sigma^2 \)

we as \( x \to x + p \) only shifts \( k \) by an additive constant \( ikp \), we only see "central" moments for coefficients of \( k^2 \) and above.

(ii) Central Limit Theorem

Let \( y = \frac{1}{n} \sum_{i=1}^{n} x_i \); \( x_i \) independent

If \( x_i \) are Gaussian then so is \( y \); in fact if \( y \) is Gaussian exactly then so must all the \( x_i \) be Gaussian.

(2) De Moivre in 1733 showed that if \( x_i \) binomially distributed, then \( y \) was always Gaussian for large \( n \) (even though Gauss did not exist in 1733).
Laplace, in 1812, showed that de Moivre's result was generally true.

**Central Limit Theorem of Laplace**

If $x_1, \ldots, x_n$ are independent random variables with the same distribution, then $y$ is asymptotically Gaussian with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$ (where $x_i$ has mean $\mu$ and s.d. $\sigma_i$, and both $\mu$ and $\sigma$ exist i.e. are finite).

**Central Limit Theorem - Liapunoff**

If $x_1, \ldots, x_n$ are independent random variables with $\mu_i = \langle x_i \rangle$

$$\sigma_i^2 = \langle (x_i - \mu_i)^2 \rangle$$

$$s_i^3 = \langle |x_i - \mu_i|^3 \rangle$$

where $s_i$ must exist (not necessarily in Laplace's sense, where all $x_i$ had the same distribution). Let

$$s^3 = \sum_{i=1}^{n} s_i^3$$
\[ \sum_{i=1}^{n} \sigma_i^2 \]

Then if \( \lim_{n \to \infty} \frac{S}{\Sigma} \to 0 \), then \( S \) is Gaussianly distributed with mean \( \mu = \sum_{i=1}^{n} \mu_i / n \) and s.d. \( \Sigma / n \).

Note the \( \lim_{n \to \infty} \frac{S}{\Sigma} \) does \( \to 0 \) when \( S \) exists in Laplace's case as \( \rho \propto n^{-\frac{1}{3}} \) and \( \Sigma \propto n^{-\frac{1}{2}} \).
Last week we stated the central limit theorem. Now we must prove one version of it.

**Proof (Chaplace):**

From our discussion of characteristic functions (c.f.),

\[ \varphi_Y = \left[ \varphi_X \left( \frac{k}{n} \right) \right]^n \]

where \( \varphi_Y / \varphi_X \) are c.f. of \( Y \) and \( X \) respectively.

Now \( \varphi_X(k) = \exp(ikp) \int dx e^{ik(x-x')} \rho(x) \)

\[ = \exp(ikp) \left[ 1 - \frac{b^2x^2}{2} + O(b^3) \right] \]
so \( \psi_y = \exp(ikp) \left[ 1 - \frac{k^2 \sigma^2}{2n^2} + O\left(\frac{1}{n^3}\right)\right]^{\frac{n}{k}} \)

\[ = e^{ikp} \left\{ \exp \left[ -\frac{k^2 \sigma^2}{2n^2} + O\left(\frac{1}{n^3}\right)\right] \right\}^{\frac{n}{k}} \]

\[ = \exp(ikp) \exp \left\{ -\frac{k^2 \sigma^2}{2n^2} + O\left(\frac{1}{n^2}\right)\right\} \]

\( \approx 0 \) as \( n \to \infty \)

and Fourier transform of a Gaussian, is well known to be a Gaussian, so

\[ p(y) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\} \]

and \( y \) is Gaussianly distributed as claimed.

(iii) **Monte Carlo Integration**

Let's consider this important numerical technique from a formal point of view. We want to evaluate

\[ I = \int_a^b f(x) \, dx \]

we can surely write:
\[
I = \int_{-\infty}^{\infty} f(x) p(x) \, dx
\]

where \( p(x) = \frac{1}{(b-a)} \quad a \leq x \leq b \)

\( = 0 \quad \text{outside} \quad [a,b]. \)

let \( \bar{I} = \frac{1}{n} \sum_{i=1}^{n} \tilde{f}(\tilde{x}_i) \)

where \( \tilde{x}_i \) are distributed according to \( p(x) \). Now apply central limit theorem to \( \tilde{x}_i = f(x_i) \).

\( \tilde{x}_i \) has mean \( \mu = \int_{-\infty}^{\infty} f(x) p(x) \, dx \)

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

Then according to theorem

\[
\bar{I} \to I \pm \sigma/\sqrt{n}
\]

or \( I \) can be estimated as \( I \)

with error \( \sigma/\sqrt{n} \). Clearly a good

made cars evaluation requires small \( \sigma \).

(i.e. small "variance" of \( f \)). We will

return to this in numerical methods.
Note that in any Monte Carlo calculation one should calculate energies as well as value of integral. It involves \( \int_a^b f^2(x) p(x) dx \)

\[
\Rightarrow \langle f^2(x) \rangle
\]

which is estimated as

\[
\frac{1}{n} \sum_{i=1}^{n} f^2(x_i)
\]

So one should accumulate \( \Sigma f^2 \) as well as \( \Sigma f \).
(iv) Extension of Central Limit Theorem

An important extension of the central limit theorem (necessary when discussing moments) is

Theorem

Let \( x_i \) be independent random variables with the same distribution and let \( f(x), g(x) \) be any functions over this distribution. Suppose

\[
\begin{align*}
\mu_f &= \langle f \rangle \\
\mu_g &= \langle g \rangle \\
\sigma_f^2 &= \langle (f - \mu_f)^2 \rangle \\
\sigma_{fg} &= \langle (f - \mu_f)(g - \mu_g) \rangle \\
\sigma_g^2 &= \langle (g - \mu_g)^2 \rangle \\
3 &= \begin{bmatrix} \sigma_f^2 & \sigma_{fg} \\ \sigma_{fg} & \sigma_g^2 \end{bmatrix} \text{ moment matrix}
\end{align*}
\]

Finally set

\[
F = \frac{1}{n} \sum_{i=1}^{n} f(x_i)
\]

\[
G = \frac{1}{n} \sum_{i=1}^{n} g(x_i)
\]
Then \((F,G)\) is asymptotically a (2-dimensional) Gaussian distribution with

\[
p(F,G) \propto \exp \left\{ -\frac{n}{2} (y - \mu)^T B (y - \mu) \right\}
\]

where we have lapsed into vector notation: \(y = \begin{bmatrix} F \\ G \end{bmatrix}\), \(\mu = \begin{bmatrix} \mu_F \\ \mu_G \end{bmatrix}\) and \(B^{-1} = M\).

**Proof:** just as in ordinary central limit theorem except you consider 2-dimensional characteristic function \(\varphi(k_1, k_2)\).

**Corollary:** if \(F, G\) are defined as above, then \(F/G\) is asymptotically Gaussian with mean \(\mu_F/\mu_G\) and standard deviation

\[
\sqrt{\frac{1}{n} \left< \frac{1}{\mu_G} \left< \frac{1}{\mu_G} \right. - \frac{1}{\mu_F} \right)^2}.
\]

**Examples**

(a) \(G = 1\), mean = \(\mu_F\), \(\sigma = \sqrt{\frac{1}{n} \sigma_F^2}\)

i.e. ordinary central limit theorem

(b) \(F = G\), mean = 1, \(\sigma = 0\).
This extension is important as it allows you to use correlations between functions calculated from the same data (observations).
Example

1) Suppose we select a sample of events from some set e.g. the sample may be those falling into some bin of a histogram. We wish to estimate the fraction of events with property $\theta$.

Define random variable $\theta$ by $\theta = 1$ event passes selection criterion $\theta = 0$ event fails criterion.

Then $f = \theta \quad <f> = p$, say.

$g = 1. \quad <g> = 1$

The estimate of fraction is just $\frac{\sum_{i=1}^{n} \theta_i}{\sum_{i=1}^{n} 1}$. 
Then error in estimate is

\[ \sigma = \frac{P}{\sqrt{N}} \sqrt{\left< \frac{f}{p} - 1 \right>^2} \]

\[ = \frac{P}{\sqrt{N}} \sqrt{\left< \frac{f^2}{p^2} - 2 \frac{f}{p} + 1 \right>} \]

\[ = \sqrt{\frac{P(1-p)}{N}} \quad \text{as } f^2 = f. \]

This is (familiar) formula for standard deviation of a binomial distribution.

The naive estimate is

\[ <f> = \frac{1}{n} \sum_{l=1}^{n} f(x_i) \]

\[ <f^2> = \frac{1}{n} \sum_{l=1}^{n} f^2(x_i) = <f> \]

\[ \sigma = <f^2> - <f^2> = p - p^2 \]

error is \( \frac{\sigma}{\sqrt{N}} \) in agreement with above.
ii) Another example:

Suppose a scattering experiment observes $\Theta$ with probability $p(\Theta)$

Then typically, one would expand

$$p(\Theta) = \sum_{n=0}^{\infty} a_n P_n(\cos \Theta)$$

as a series of Legendre Polynomials.

Then

$$\langle P_{n_1}(\cos \Theta) \rangle = \int P_{n_1}(\cos \Theta) p(\Theta) \, d\Theta$$

$$= a_{n_1} \frac{2}{2n_1 + 1}$$

$$\therefore \langle P_{n_1}(\cos \Theta) \rangle \text{ need to be calculated as averages over observed } \Theta_k$$

$$\langle P_{n_1}(\cos \Theta) \rangle = \frac{1}{N} \sum_{k=1}^{N} P_{n_1}(\cos \Theta_k)$$

Even though $P_{n_1}(\cos \Theta)$ and $P_{n_2}(\cos \Theta)$ are orthogonal polynomials, the means $\langle P_{n_1}(\cos \Theta) \rangle$ and $\langle P_{n_2}(\cos \Theta) \rangle$ are not uncorrelated. Both
are calculated from same set \( \{\theta_k\} \) and fluctuate in a correlated fashion as observables \( \theta_k \) fluctuate.

\[
\begin{align*}
f &= P_{n_1} (\cos \theta) \\
g &= P_{n_2} (\cos \theta) \\
F &= \frac{1}{N} \sum_k P_{n_1} (\cos \theta_k) \\
G &= \frac{1}{N} \sum_k P_{n_2} (\cos \theta_k)
\end{align*}
\]

\( F \) and \( G \) will be independent if

\[
\int p(\theta) P_{n_1} (\cos \theta) P_{n_2} (\cos \theta) = 0
\]

This is not in general true.
iii) Consider a histogram in which \( N \) events with weight \( w_i \) \( (i=1...N) \) and \( x \)-value \( x_i \) fall.

\[
\langle x \rangle = \frac{\sum_{i=1}^{N} w_i x_i = S_3}{\sum_{i=1}^{N} w_i = S_4}
\]

\[
\text{c.f. Monte Carlo calculation of } \int x \omega(x) dx / \int \omega(x) dx
\]

with \( f = \omega x \) \( g = \omega \)

the error in this calculation is:

\[
\sigma = \frac{S_3}{S_4} \sqrt{\frac{1}{N} \left\langle \left( \frac{N \omega x}{S_3} - \frac{N \omega}{S_4} \right)^2 \right\rangle}
\]

\[
\sigma^2 = \frac{S_3^2}{S_4^2} N \left\langle \left[ \frac{\omega^2 x^2}{S_3^2} - \frac{2 \omega^2 x}{S_3 S_1} + \frac{\omega^2}{S_1^2} \right] \right\rangle
\]

\[
= \frac{1}{S_1^2} \left[ S_6 - 2 \langle x \rangle S_4 + \langle x^2 \rangle S_2 \right]
\]

\[
S_2 = \sum_{i=1}^{N} w_i^2
\]

\[
S_4 = \sum_{i=1}^{N} w_i x_i
\]

\[
S_6 = \sum_{i=1}^{N} w_i x_i^2
\]
D3. Seafood in Las Vegas

D3/1: Bernoulli and Poisson Distributions

we now come to change of space and consider more combinatorial aspects of probability. Much the nearest reference for this is Feller who has a host of beautiful examples.

(i) Bernoulli Trials

(ii) A series of Bernoulli Trials is the generalized coin tossing problem. We have a sequence of independent "trials" - each of which has only two outcomes: success or "heads"; probability \( p \);

failure or "tails"; probability \( q \);

where of course \( p + q = 1 \).

Efficienators of affirmative action and women's lib, will note that
generations of math books have downgraded tails compared to heads.

Anyhow we follow in their footsteps and set \( b(k, n, p) \) as the probability of \( k \) successes in \( n \) trials. This is easily calculated by writing all possible outcomes as, for instance,

\[
\underbrace{hthttt\ldots}_{n}
\]

There are \( \binom{n}{k} \) configurations with \( k \) "h"'s and each has probability \( p^k q^{n-k} \). So:

\[
b(k, n, p) = \binom{n}{k} p^k q^{n-k} \quad (D3.1)
\]

This is the binomial distribution because \( b(k, n, p) \) are just terms in binomial expansion

\[
1 = (p+q)^n = \sum_{k=0}^{n} b(k, n, p).
\]
(b) Example:

A course consists of 50 lectures given to 20 people—all of whom get up before 9 a.m.—and attend all classes. A grand total of 1000 smiles is observed. What is the chance that Frank Nagy smiled exactly once during the whole course.

Self: Use previous formalism, setting each trial as a smile; and success as Frank Nagy smiled. A priori each smile has equal probability of belonging to each person.

So \( p = \frac{1}{20} \), \( q = \frac{19}{20} \), and \( n = 1000 \).

Thus:\n
\[ \Pr \{ \text{Frank Nagy smiles once and once only} \} = \]
\[
\binom{1000}{1} \cdot \frac{1}{20} \cdot \left(\frac{19}{20}\right)^{999} = 50 e^{999 \log_2 19/20} \]

which isn't very big

\[
\approx 50 e^{-51}
\]

\[
\approx 10^{-19}
\]

\[\text{(ii) Poisson limit:}\]

(a) This is a very important special case of (23.3) in the binomial distribution (33.1) which describes anything from telephone calls to events observed in particle physics. In (33.1), take limit of \(n \to \infty\) and \(p \to 0\) such that \(np = \lambda\) remains finite. \(k\) is also fixed.

\[
\binom{n}{k} = \frac{n!}{k! (n-k)!} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k}
\]

now

\[
\frac{n!}{(n-k)!} \to n(n-1) \cdots (n-k+1) \to n^k\text{ as } n \to \infty\text{ for fixed } k.\] (you can also get this using Stirling's formula for the
Gamma function

also \((1 - \frac{1}{n})^{n-k} \approx (1 - \frac{1}{n})^n \rightarrow e^{-\lambda}\)

so: \(p(k, n, \lambda) \rightarrow \frac{n^k}{k!} \frac{x^k}{n^k} e^{-\lambda}\)

\[ = p(k, \lambda) = e^{-\lambda} \frac{x^k}{k!} \quad (D3.2)\]

→ The Poisson distribution.

Note \(\sum_{k=0}^{\infty} p(k, \lambda) = 1\) as it should

Example: Given circumstances of a previous example, what is chance that Frank Nagy smiles exactly once in a given lecture?

Sott: Again we set each trial = a smile and we have a 1000 of them. Now success = "Frank Nagy smiled this smile in our given lecture" and it is not very probable \(p = \frac{1}{20.50} = \frac{1}{1000}\).
So $n$ is large, $p$ is small but $np = \lambda = 1$ is reasonable.

So the Poisson approximation to the exact (in this example) Binomial distribution is valid and chance is $e^{-1}$ i.e. quite probable that F.N. Smiles once and once only in a given lecture.

The Binomial estimate is

$$\binom{1000}{1} \frac{1}{1000} \left(\frac{999}{1000}\right)^{999} = \left[\frac{999}{1000}\right]^{999}$$

which is indeed $\approx e^{-1}$. 
Compound Poisson Distribution

\[ y = \sum_{i=1}^{N} y_i \]

\( y_i \) is Poisson with mean \( \mu \)

\( N \) is Poisson with mean \( \lambda \)

\[ \Pr (y = \text{integer} r) \text{ is} \]

\[ \sum_{N=0}^{\infty} \Pr (N \text{ takes given value}) \cdot \Pr (\sum_{i=1}^{N} y_i, N \text{ fixed has } r \text{ events}) \]

As for fixed \( N \), \( y \) is Poisson with mean \( N \mu \), we find

\[ \Pr (y = \text{integer} r) \]

\[ = \sum_{r=0}^{\infty} \frac{\lambda^N e^{-\lambda}}{\lambda^r} \cdot \frac{(N \mu)^r e^{-N \mu}}{r!} \]

More generally, we could give \( y_i \) a non-Poisson distribution.

Examples are
1) Ionization - p 53 EDTRE

Particularly interesting is sensitivity to assumption that Poisson
note that \( N \) variables with each
with note the Poisson distribution of
4 objects with \(< \text{number of objects} > = \lambda \)
is much broader than ordinary
Poisson of mean 4\( \lambda \). This is of
importance in particle physics where
ii) we believe that \( q\bar{q} \) "clusters"
are produced independently in a
high energy collision. These clusters
are Poissonly distributed and decay
with \( 4 \to 6 \) final particles.

Show \( f_2 = \sigma^2 - \text{mean} = 0 \) for Poisson
\( > 0 \) for cluster model.

This is a compound process with
\( y_i \) probably distributed Gaussianly.
For example, consider Poisson distribution of clusters—each of which always contains 4 particles. For a Poisson distribution

\[ \text{Mean} = \lambda \quad \text{Var} = \lambda \]

\[ f_2(x) = \sigma_x^2 - <x> = 0 \]

let random variable \( y = 4x \)

\[ <y> = 4 <x> \]
\[ <y^2> = <(4x)^2> = 16 <x^2> \]
\[ = 16 (<x^2> - <x>^2 + <x>) \]
\[ = 16 (\sigma_x^2 + <x^2>) \]
\[ = 16 (\lambda + \lambda^2) \]

\[ f_2(y) = <y^2> - <y>^2 \]
\[ = 16 (\lambda + \lambda^2) - 16\lambda - 4\lambda \]
\[ = 12 \lambda > 0 \]

\[ \therefore y \text{ is always inside than } x \]
Cherenkov light
Total pulse height observed

\[ y = \sum_{i=1}^{N_x} y_i \]

\[ y_i = \text{response in phototube of a single photon - probably Gaussian} \]
(i.e. limit of large Poisson)

\[ N_x = \text{number of photons given off in Cherenkov medium which is certainly Poisson.} \]

\[ P(y) \]

\[ \text{discernible photon electron peaks} \]

\[ \text{Poisson smeared by Gaussian.} \]
(iii) Additivity Property of Binomial or Poisson Distributions

These two distributions share an important additivity property which (is a) obvious and (b) leads to confusion when used in conjunction with central
limit theorem.

(a) Suppose the independent random variables \( X_1 \) and \( X_2 \) have binomial distributions with parameters \((n_1, p)\) and \((n_2, p)\) respectively. Here \( X \equiv k \) if \( k \) successes. Then it is obvious from source of binomial distribution that \( X_1 + X_2 \) has a binomial distribution with parameter \((n_1 + n_2, p)\).

(b) Suppose independent random variables \( Y_1 \) and \( Y_2 \) have Poisson distributions \( \lambda_1 \) and \( \lambda_2 \). Then \( Y_1 + Y_2 \) has Poisson distribution with parameter \( \lambda_1 + \lambda_2 \). This is obvious either algebraically or by using (a) with \( n_1 = \lambda_1 / p \) and \( n_2 = \lambda_2 / p \), and let \( n_i \to \infty, p \to 0 \), fixed \( \lambda_1, \lambda_2 \).
(iv) Gaussian limit of Binomial Distribution

(a) m. and n. confused me by discussing two limits of binomial distribution.

We have already discussed one of these - the Poisson limit - which is really distinctive limit.

However we can also see a Gaussian limit which is just a special case of central limit theorem. Although this limit has a particular interpretation because of the additivity property we just discussed. Note also that central limit theorem was discovered in general in 1812 whilst de Moivre discovered its binomial special case in 1733.

(b) let $x_i$ be independent random
variables - each of which has the same \( b(k, 1, p) \) distribution, i.e., for each \( i : x_i = 0 \) is failure and \( x_i = 1 \) is success with relative probabilities \( 1-p \) and \( p \) respectively. Then \( y = \sum_{i=1}^{k} x_i \) is in fact \( \text{(B)} \), the number of successes in \( n \) Bernoulli trials. Then

(a) from additivity \( \text{(iii)} \): \( y \) is Binomial with distribution \( b(k, n, p) \)

(b) From Standard central limit theorem, \( y \) is Gaussian with mean \( np \) and standard deviation \( \sqrt{n} \). Here \( p \) and \( n \) are mean and S.D. of \( b(k, 1, p) \) distribution, i.e.,

\[
\mu = 0 \cdot (1-p) + 1 \cdot p = p
\]

\[
\sigma^2 = < x^2 > - < x >^2 = 0 \cdot (1-p) + 1 \cdot p - p^2 = p(1-p)
\]
We conclude that for fixed $p^n$, and $n \to \infty$, $b(k, n; p)$ tends to Gaussian distribution with mean $np$ and standard deviation $\sqrt{p(1-p)n}$.

(c) We can do the same trick for the Poisson distribution. Fix $\lambda$ finite and let $x_i$ be independent with each as Poisson with parameter $\lambda$. Then from additivity $y = x_1 + \ldots + x_n$ has Poisson distribution with parameter $n \lambda = \lambda$. Meanwhile from central limit theorem, $y$ is Gaussian with mean $np$ and standard deviation $\sqrt{n \sigma}$. Here again $p$ and $\sigma$ are mean and s.d. of individual distribution.

\[
p = \sum \exp(-\lambda) \frac{\lambda^k}{k!} = \lambda \int_0^\infty e^{-\lambda} \varphi (\lambda) = \lambda e^{-\lambda} \frac{d}{d\lambda} (e^{\lambda}) = \lambda.
\]
\[ \sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 \\
= e^{-\lambda} \left( \lambda \frac{\partial}{\partial \lambda} \right)^2 (e^\lambda) - \lambda^2 \\
= e^{-\lambda} \left( \lambda \frac{\partial}{\partial \lambda} \right) (\lambda e^\lambda) - \lambda^2 \\
= \lambda + \lambda^2 - \lambda^2 = \lambda = \mu \]

So, for large \( \lambda \), any Poisson-distributed random variable \( y \) has mean \( \lambda \) and standard deviation \( \sqrt{\lambda} \). We will soon show that any counting experiment (e.g., number of raindrops in a bowl or particles in a high energy physics experiment) is described by a Poisson random variable with (large) parameter \( \lambda = \text{number of expected counts} \). The above result then translates into well-known result that counting experiment with \( N \) events, has error \( \sqrt{N} \).
D3/2 Birth and Death Processes

(a) We will briefly sketch some techniques which at their most elementary justify Poisson for a counting experiment. The more advanced applications are important in many practical problems. (Telephones (random calls) and Biology (random birth/deaths)) involving happenings distributed more or less randomly in time.

(b) Telephone calls

We start with a discrete example:

Suppose your favourite secretary receives $N$ telephone calls in an hour. We wish to find probability that she receives $k$ calls in a given time.
Consider every second; if probability
\[ p = \frac{N}{3600} \]
of receiving one call; if essentially no probability of receiving
>1 call. Thus each second can be
described by variable \( y_i \) (\( =0 \) no, \( =1 \) one call) which has Poisson distribution
with parameter \( \lambda = \frac{N}{3600} \). The
number of calls in time \( t \) seconds
is \( \sum_{i=1}^{t} y_i \) which from additivity is
also Poisson with parameter \( \sum_{i=1}^{t} \lambda_i = pt \). So
\[ P_k(t) = e^{-pt} \frac{(pt)^k}{k!} \]
is probability
of receiving \( k \) calls in \( t \) seconds.
Note mean number received is \( pt \)
or for \( t = 3600 \), just \( N \) as input
data claimed.

One could also derive this
by taking each second as a Bernoulli trial with probability $p$ of success then we get distribution $b(k, t, p)$ which has $t \to \infty, p \to 0$ pt finite i.e. we can employ usual Poisson limit.

(c) Basic Birth Process

It is natural to take limit of time interval (second in above example) to zero and let $t$ be a non-continuous and not a discrete valued variable. Suppose we are studying events (e.g. telephone calls as above) that

1). Occur independently of previous history of world.

2). In time $[t, t+dt]$, the probability of an event happening is $\lambda dt + o(dt)$ and
the probability of \( >1 \) event happening is \( o(dt) \).

Define \( P_k(t) \) as the probability of \( k \) events happening in time \( 0 \leq t \). Then

\[
P_k(t+dt) = P_k(t) \left[ 1 - \lambda dt \right]
\]

i.e. \( k \) events

\[
= P_{k-1}(t) \lambda dt + o(dt)
\]
at \( t+dt \)

or

\[
\frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t) \quad : k \geq 0
\]

For \( k=0 \), the above is modified as

\( P_{-1} \) item doesn't exist i.e.

\[
\frac{dP_0(t)}{dt} = -\lambda P_0(t) \quad : k=0
\]

Solving the last gives

\[
P_0(t) = e^{-\lambda t} \quad \text{and then induction}
\]

Shoos \( P_k(t) = e^{-\lambda t} (\lambda t)^k / k! \)

and we have achieved Poisson again

It is now obvious why any counting experiments for random events
should have a Poisson distribution.

(d) **Birth and Death Process (Feller p. 40)**

Consider a system with a denumerable number of state \( E_n \) whose probability of occupation \( P(\cdot) \) varies with \( t \). For Poisson \( E_n \), it is "we have observed \( n \) events" (not only one \( E_n \) occupied at a time) and \( \frac{d}{dt} \) change \( \lambda dt \) of a transition \( E_n \to E_{n+1} \). This is a pure birth process.

We can generalize \( \lambda \) to \( \lambda_n(t) \): a function of \( n \) and \( t \). Also we have \( \lambda \) death processes which also allow relapses \( E_{n-1} \to E_{n+1} \) with chance \( \mu_n(t) dt \). So:

- chance \( E_n \to E_{n+1} \)
- chance \( E_n \to E_{n-1} \)
- Any other change: \( \mu_n(t) dt + o(dt) \)

We can at once, derive a similar
differential equation for \( P_n \): probability of
system being in state \( E_n \).
\[ \frac{dP_n(t)}{dt} = - (\lambda_n + \mu_n) P_n(t) \]
\[ + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t) \]
with \( P_0 = 0 \) and \( \lambda_{-1} = 0 \) so that alive
is still counted with \( n=0 \).

(e) Example of Birth/Death Process

Consider a telephone exchange with
an infinite number of lines and let
\( E_n \) be state "\( n \) lines are busy". Then
we apply alive formalism with
\[ P_n = n \mu \quad \mu \text{ fixed.} \quad (\frac{\lambda}{\mu} \text{ is average length of call}), \]
but \( \lambda_n = \lambda \) independent of \( n \).

One can find the steady state
solution independent of both \( t \) and
initial state:
\[ P_n(t) = e^{-\lambda/\mu} (\lambda/\mu)^n / n! \quad \text{as} \]
probability of \( n \) lines being busy.
It is also easy to do the seemingly harder problem of a finite number of telephone lines per exchange. Here is an example of this sort in problem set about parking lots.

Similar techniques can be used to decide on number of servicemen needed to repair a set of machines of prescribed reliability. Here

\[
\text{Death} = \text{machine starts working} \\
\text{Birth} = \text{machine breaks down.}
\]

and \( E_1 = n \) machines broken. There is a beautiful discussion of this on pages 416-420 of Feller.

(g) Markov Chains

One can generalize either the discrete (case (b)) or continuous (cc) to (c) cases by considering a system
with again states $E_n$ and now not only transitions $E_n \to E_{n+1}$ but a general probability $P(n \to k)$ or $P(k \to n)$ of transition from $E_j$ to $E_k$ at each trial.

This theory includes general random walk problem and is discussed in Vol. 1, Feller. The mathematics is not very deep—solving recurrence relations—and the techniques are not useful for the basic problem in physics (Statistics of data reduction). So I will not consider further. People with biological bent should however read further in Feller.

This Markov theory is relevant in discussions of the so-called Metropolis method in Monte Carlo integration.
These notes refer to some additional references:

Cramer: Mathematical Methods of Statistics
published by Princeton.

plus

and some old notes I have copies of

J. Orear, "Notes on Statistics for Physicists"
UCRL-8417 (1953)

S. Yellin, DESY 72/12: Preprint (1972)
K. McDonald CTSL Internal Report # 57.
Estimation of Parameters

So far we have been concerned with rather straightforward problems on random variables with known probability distributions. Now we turn to a fundamental scientific problem: given some observations $x_1, \ldots, x_n$ (= values of a random variable) of some physical quantity, how do we extract from the theoretical parameters describing the system we have observed and how do we estimate error in our determination. This will require rather more careful thought as to relation of probability theory and real world. We will first describe the principle of maximum likelihood - everything else can be derived from this.
D4/1 Maximum Likelihood Principle

(i) Example:

(a) Rod Shivering in an Earthquake

\[ x_1, x_2, \ldots, x_n \] are \( n \) measurements of the length of a rod. Owing to a Gaussian earthquake (or a Gaussian [i.e. normal] darkness) occurring during measurement, the \( x_i \) are scattered about the true value. You visit your favourite theonist (or do yourself a favour if solving up) who believes that the probability distribution of \( x_i \) is normal with mean \( \mu_0 = \text{true length} \) and standard deviation \( \sigma_0 \).

We may write down the (relative) probability of each measurement \( x_i \) as:

\[ p_i(x_i) = c \exp \left\{ -\frac{1}{2} \sigma_0^2 (x_i - \mu_0)^2 \right\} \]
The likelihood function for whole experiment of \(n\) events is defined as:

\[
\mathcal{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{\exp \left\{ -\frac{1}{2\sigma_0^2} (x_i - \theta_0)^2 \right\}}
\]

\[= e^{\exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i} (x_i - \theta_0)^2 \right\}}\]

This is - up to a constant - the probability of the \(x_i\) given the true length \(\theta_0\) and certain theoretical assumptions we represent as \(T\).

\[
\mathcal{L}(x_1, \ldots, x_n) = \Pr\{x_1, \ldots, x_n \mid \theta_0, T\}
\]

in conditional probability notation. \(T\) is for instance the notion that dislodgment Gaussian and the value of \(\theta_0\) taken (the latter being calculable from the strength of earthquake or amount of bear drink).

\((b)\) Enter Mr. Bayes
Now what we really want is the probability distribution of an estimate $x$ of $x_0$. We can formally get this by use of Bayes' law: first extend $p(x_i) = \exp \left\{ -\frac{1}{2} \sigma^2 (x_i - x)^2 \right\}$ to be defined for any value of $x_i$ not necessarily $x = \text{true value } x_0$. Introduce:

$$P(x \mid x_i, T) = \frac{P(x_i \mid x, T) P(x \mid T)}{P(x_i \mid T)} \quad (D4.1)$$

which only differs from ordinary Bayes by extra IT in each probability.

This is the essence of the maximum likelihood principle. \( (D4.1) \) converts the likelihood $L(x_1, x_n) = P(x_i \mid x, T)$ which is a function of $x_i$ for fixed
\( \alpha, T \) into \( P(\alpha | x_i, T) \) which is the desired probability distribution of \( \alpha \) for given measurements \( x_i \) and prejudice.

Interpretation of (B4.1)

(a) We can ignore \( P(x_i | T) \) because it is independent of \( \alpha \) and fixed for our given \( x_i \) and \( T \). It has the practical point that likelihood is never normalized sensibly.

(b) \( P(x_i | T) \) philosophically represents a prior knowledge of \( \alpha \) before experiment (measurement of \( x_i \) performed). This is shaky - not because we don't really know how to define what it ("given \( T \)"") means - but, for instance, suppose we wish to parameterize "complete ignorance" of \( \alpha \) before
experinsl perfonmz.

Natural is $P(x|T) = \text{constant}$

But suppose we put $\beta = x^2$ and then $p$ has a prior probability

$$q(\beta|T) \, d\beta = P(x|T) \, dx$$

i.e. $q(\beta|T) = \frac{1}{2\sqrt{\beta}} \neq \text{constant}$

So the concept of "complete ignorance" is not well defined. Jefiferz (a big bad English-oxford-book cited in mathen and wather) spends 100's of pages hving to avoid this difficulty. His efforts are so much garbage.

Fortunately there is no trouble for large $n$ because of both an obvious numerical reason (see next Section (i)(a)) and a not so obvious mathematical reason (see Section (ii)).
The likelihood maximizes

Consider, again, the explicit form of the likelihood of a rod quivering in the earthquake.

\[ l(x_1, \ldots, x_n) = c \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \]

According to (D4.1), we now regard \( l \) as a function of \( \mu \) and so rewrite \( l \):

\[ l(\mu) = c \exp \left\{ -\frac{n}{2\sigma^2} (\mu - \frac{1}{n} \sum_{i=1}^{n} x_i)^2 \right\} \]

\[ -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2 \right] \]

The last item is a constant independent of \( \mu \), and so as a function of \( \mu \), \( l \) is just a Gaussian which maximizes at \( \mu = \frac{1}{n} \sum_{i=1}^{n} x_i \) and has width \( \sigma \sqrt{n} \).
So as \( L = P_r(\alpha | x_i, T) \) we say that maximum likelihood estimate of \( \alpha \) is \( \hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} x_i \) and its standard deviation is \( \sigma / \sqrt{n} \).

Note that whatever \( P(\alpha) \) you put in the shape of \( L \cdot P(\alpha | T) \) is essentially independent of \( P \) as now and we get a Gaussian peaked at \( \hat{\alpha} \) riding on top of constant value \( P(\alpha | T) \).

(ii) General maximum likelihood method

(a) The above is glib in many ways—especially in its treatment of errors. We will repair this later with real live theorems. First we state the obvious generalization.

The general maximum likelihood method is to put \( P(\alpha | T) = 1 \) in (D4.1) and form \( L(\alpha) = \prod_{i=1}^{n} P(x_i, \alpha) \) for any
measured $x_i$ whose probability $p(x_i, \alpha)$ depends on value $x_i$ and theoretical parameter(s) $\alpha$.

The maximum likelihood estimate of $\alpha$ is $\hat{\alpha} = \alpha^*$ where $L(\alpha^*)$ is a maximum. The error is the standard deviation of the distribution of $L$ as a function of $\alpha$, i.e.

$$L(\alpha) = \exp \left\{ \text{const} - \frac{1}{2} \Delta \alpha^2 (\alpha - \alpha^*)^2 \right\}$$

(b) One must issue some obvious warnings. For instance $L$ could have several maxima.

![Graph showing multiple maxima]

For instance, this would happen if in earthquake problem you were still drunk in morning and decided to
find $\tilde{F}$. Then $\tilde{F}$ would peak at $\pm \tilde{F}_0$.

less obviously one should remember that theorems (to come) only say that maximum likelihood method works for large $n$. Then it is a universally good method. However, for small $n$ it is much more likely to give nonsense than many other methods. Thus...

(c) Warning Example

Consider probability distribution appropriate for galvanometer exp

$$p(\varphi, \alpha) = \frac{1 + \alpha \cos \varphi}{2\pi}, \quad -1 \leq \alpha \leq 1.$$  

If we had but one event

$$\alpha (\alpha) = \frac{1 + \alpha \cos \varphi_1}{2\pi}$$

if $\cos \varphi_1 > 0 : \alpha = 1$ (i.e. end of range) maximizes $d$

or $\cos \varphi_1 < 0 : \alpha = -1$ (other end of range) maximizes $d$.  
If you have 2 events:

\[ L(x) = (1 + \alpha (\cos \phi_1 + \cos \phi_2) + \alpha^2 \cos \phi_1 \cos \phi_2) \frac{1}{2 \pi} \]

This has a maximum at end of range, again, unless \( \cos \phi_1, \cos \phi_2 < 0 \).

(d) So usually people warn you by saying: "If distribution of \( \alpha \) is not Gaussian, then do not simply quote maximum and standard deviation, rather plot the likelihood". Unfortunately this is dubious, because for small \( n \), nobody has proved any theorems and \( P(\alpha|T) \) philosophy important while for large \( n \), theorems are proven on the relevance of the likelihood method but they also prove \( L \) is Gaussian!

So correct warning is: "If
distinguish not Gaussian - everything is OK if either:

(a) just many sharp humps as in \( T \) example, data can be fitted with >1 value of parameter.

(b) if several parameters, can be Gaussian in some and flat in others if depends weakly (or not at all!) on them.

(c) Gaussians maybe distorted by effects d due to end of range.

if (a) (b) or (c) don't apply and \( x \) is not Gaussian then one can only be unhappy!
(iii) lifetime Example

All books give the example of determining the lifetime $\tau_0$ of a particle from the observation of decay times $t_1, \ldots, t_n$ of $n$ decays.

The probability distribution is Poisson:

$$p(t) \, dt = \frac{1}{\tau} \exp \left( -\frac{t}{\tau} \right) \, dt$$

Now $\, dt$ is meaningless - it is independent of $\tau$ and only affects normalization of likelihood. The unknown $P(x|\tau), P(x;|\tau)$ terms in (A4.1) made this arbitrary anyway!

Likelihood $L(\tau) \propto \frac{1}{\tau^2} e^{-t_1/\tau} \cdots \frac{1}{\tau} e^{-t_n/\tau}$

or

$$L(\tau) = \exp \left[ -\frac{1}{\tau} \sum_{i=1}^{n} t_i - n \ln \tau \right]$$

where we lazily convert $\ln$ into $\ln = \text{Sign}$.
maximum is \( \frac{dL}{dz} = 0 \) or
\[
\frac{1}{\tau^2} \sum_{i=1}^{n} b_i - \frac{n}{\tau^2} = 0
\]
or \( \tau^2 = \frac{1}{n} \sum_{i=1}^{n} b_i \) or the maximum likelihood estimate is just the mean.

Such a simple result is only true because of particular Poisson example.

We can use this example to illustrate an interesting theoretical point about the error.

(a) According to the central limit theorem, the variable \( y = \frac{1}{n} \sum_{i=2}^{n} \) is a sum over independent random variables and has mean \( \langle y \rangle = \tau_0 \) and standard deviation \( \sigma \)

\[
\sigma^2 = \frac{1}{n} \left\{ \langle b_i^2 \rangle - \langle b_i \rangle^2 \right\}
\]

\[
= \frac{1}{n} \left\{ \frac{1}{\tau_0} \int t^2 e^{-t/\tau_0} dt - \tau_0^2 \right\}
\]

or \( \sigma = \frac{\tau_0}{\sqrt{n}} \) (D4.2)
(vi) Multiplication of Experiments

If we have two experiments that are independent but depend on same theoretical parameter $\alpha$.

Then $L_1 = \prod_{i=1}^{n_1} p(x_i; \alpha)$ is likelihood of exp #1

$L_2 = \prod_{i=1}^{n_2} q(y_i; \alpha)$ is likelihood of exp #2

and clearly the combined likelihood is

$L = L_1 L_2$

i.e. one should multiply likelihoods when combining experiments. We will now apply this to case

when each $L_i$ is Gaussian in $\alpha$.
\[ \chi^2 \text{ method} \]

This is standard and fully correct (best) method for combining results from several experiments. It is in fact used (because it is easy of conceptually and for computer) even when it is not applicable e.g. if there are correlations between supposed "independent" random variables or if basic experiments have too few events to be Gaussian.

The problem is: given N experimental measurements \( x_i \) and errors \( \sigma_i \) plus estimates \( \hat{f}_i \) of the \( x_i \) which are functions of \( n \) theoretical parameters. Then the probability of each observation is

\[
P_i = \frac{1}{\sqrt{2\pi} \sigma_i} \exp \left\{ -\frac{(x_i - \hat{f}_i)^2}{2\sigma_i^2} \right\}
\]
The likelihood \( L = \frac{N}{\pi} \prod_{i=1}^{N} p_i \)
\[ \propto \exp \left\{ -\sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{2 \sigma_i^2} \right\} \]

So, the principle of maximizing the likelihood is equivalent to minimizing \( \chi^2 \) where
\[ \chi^2 = \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2} \quad (D4.2.1) \]

with \( \sigma_i \) the variable parameters in \( \xi_i \).

Notice that \( \chi^2 \) is the sum of \( \frac{N}{2} \) independent Gaussian random variables of unit standard deviation and zero mean. We can calculate \( \chi^2 \) distribution exactly (m. and w. pages 393-395) but in practice this is rarely necessary. Thus, we can easily show that for fixed \( \xi_i \), that
\[ \langle X^2 \rangle = N \]
\[ \sigma_{X^2}^2 = 2N \]

and this is sufficient for large \( N \) as central limit theorem Says \( X^2 \) will be Gaussian with above mean and standard deviation.

(i) Now (D4/2.2) is not quite correct because if we determine the theoretical parameters to minimize \( X^2 \) then implicitly \( \xi_i \) are also random variables that are functions of \( X^2 \).

It is not too hard to show that in this case one should just replace \( N \) by \( N-n \) (number of degrees of freedom) in (D4/2.2). In any case, (D4/2.2) is very important because it allows an easy criterion for the goodness of fit. Thus if value
of \( x^2 \) is very (say 2) standard deviations away from its mean, then the theoretical form is in doubt. The usual maximum likelihood method does not have this advantage. Remember 1 was normalized and so its value had no significance for the goodness of fit. This is quite a difficult problem and an elementary discussion is given in Macdonald notes page 14 onwards.

(3) The only way to get a feeling for \( x^2 \) is to use it: So I want to dwell on the important techniques used to minimize \( x^2 \) to get good theoretical parameters and estimate errors in the same manner and work.
(v) Example of use of $\chi^2$

Consider the example in Mathews and Walker page 340

$$\sigma(\Theta) = a_0 + a_1 \cos \Theta + a_2 \cos^2 \Theta + a_3 \cos^3 \Theta$$

The measurements $x_i$'s are cross sections and the theoretical predictions $\xi_i$ are linear in unknown parameters $a_j$.

$$\xi_i = \sum_{m=1}^{n} c_{im} a_m$$

minimize

$$\sum_{L=1}^{N} \sum_{L=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2}$$

$$x_i = \text{const. } N_i$$

$$\sigma_i = \text{const. } \sqrt{N_i}$$

$$\sigma_i = x_i / \sqrt{N_i}$$

where $i$th bin has $N_i$ events.
\[ \sum_{i=1}^{n} \frac{(x_i - \bar{x}_i)}{\sigma_i^2} \frac{d\bar{x}_i}{d\alpha_n} = 0 \]

\[ \sum_{\ell} m_{\ell} a_\ell^* = \mathbf{B}_m \]

\[ \chi^2 \text{ max.} = \sum_i \frac{C_{im} C_{i\ell}}{\sigma_i^2} \]

Positive symmetric matrix

\[ = \sum_i \left( \frac{d\bar{x}_i}{d\alpha_n} \right) \left( \frac{d\bar{x}_i}{d\alpha_\ell} \right) \]

\[ \frac{\sigma_i^2}{\sigma_i^2} \]

\[ b_m = \sum_i \frac{x_i C_{im}}{\sigma_i^2} = \sum_i \frac{x_i}{\sigma_i^2} \frac{d\bar{x}_i}{d\alpha_n} \]

\[ a^* = m^{-1} b_m \]
(v) We can find errors in either of the following ways:

1) regard $a^*$ as a function of $x_i$ and calculate standard deviation from central limit theorem.

2) Shape of likelihood:

$$L = \exp \left(-\frac{1}{2}x^2\right)$$

we have - as far as shape of $L$ concerned

$$L = \exp \left(\text{independent of } a\right)$$

no terms linear in $a-a^*$ due to "maximum" condition

$$-\frac{1}{2} (a-a^*)^T M (a-a^*)$$

i.e. $M$ is error matrix for $a$. 
The mean and standard deviation of $\bar{y}$ in many experiments - each $n$ observations.

However, maximum likelihood method of finding error is, off hand, quite different. Rather than looking at distribution of $y$ over infinitely many estimates of $n$ observations, it considers dependence of $L$ for one given experiment of $n$ observations.

Then we found $\delta$ became Gaussian and calling maximum likelihood standard deviation $\sigma$,

$$L(\tau) \propto \exp \left\{ -\frac{1}{2\sigma^2} (\tau - \bar{y})^2 \right\}$$
or \[ \sqrt{\sigma^2} = -\frac{d^2}{dt^2} \ln L(t) \bigg|_{t=t^*} \]
\[ = \frac{2}{\tau^3} \sum_{i=1}^{\tau} t_i - \frac{n}{\tau^2} \bigg|_{t=t^*} \]

or \[ \sigma = \frac{t^*}{\sqrt{n}} \quad (D4.3) \]

as \[ t^* = t_0 + O\left(\frac{1}{n}\right) \] we see that (D4.2) and (D4.3) are in fact the same asymptotically (large \( n \), that is).

Note that physics books (Solmily is an exception) don't mention method (a) of finding error in maximum likelihood method. On the other hand, mathematics books never mention (b) as a serious method. They do however prove (a) and (b) are the same for large \( n \) (as we indicated in example above) for general distribution.
As far as I can see, method (a) is only correct definition of error. It follows that method (b) is wrong when it differs from method (a). (b) does differ from (a) for small n when, in particular, likelihood is not Gaussian, so warnings for small n ("plot likelihood") are very dubious. The correct statement is that your results are simply unreliable and error is not calculated from likelihood function alone. The error is defined but must be calculated as in (a).

Note that when math books discuss maximum likelihood estimate for small n it is only value x+ they discuss not
Shape of $L$ and method (a) of finding $\alpha$.

(v) Asymptotic Validity of Maximum Likelihood Method: Cramer chapters 22 and 23.

Theorem:

Let $P(x_i, \alpha)$ be a normalized probability for a result $x_i$ given theoretical parameter $\alpha$. We have $n$ such observations $x_i$:

$$L = \prod_{i=1}^{n} P(x_i, \alpha)$$ is usual likelihood function.

Let $\hat{\alpha}$ be maximum of $L$ w.r.t. $\alpha$.

(c) If we fix $n$ and conduct many such series of $n$ observations

$$<\hat{\alpha}> = \alpha_0 + O(1/n)$$

Note it is not $= \alpha_0$ exactly. ($\alpha_0$ is true theoretical parameter). If for any estimate $<\text{estimate}> = \alpha_0$, then we say...
estimate is unbiased: if \( \hat{\theta} = \theta_0 + b(\theta_0) \) then we have a biased estimate with bias \( b(\theta_0) \). So maximum likelihood estimate is for finite biased and is only asymptotically unbiased. However bias is smaller than standard deviation and so it is of no real importance. Bias is discussed in detail by Cramer.

(b) We can regard \( \hat{\theta} \) as a random variable: then its standard deviation is \( O(\sqrt{n}) \) and to this order we can correctly calculate standard deviation in the "physics way" already discussed.

(i.e. from a dependence of \( L(\theta) \)).

(c) There is no other estimate of \( \theta \) which has a smaller standard deviation to asymptotically in i.e. maximum
likelihood method is best as long as $n$ large.

**Proof:** (c) can be found in Cramer p. 479 or Hoel p. 379. For (a) and (b), write

$$\ln L(\alpha) = \sum_{i=1}^{n} \ln P(x_i, \alpha)$$

Expand $L$ about the theoretical values $\bar{x}$,

$$\ln L(\alpha) = L_0 + (\alpha - \bar{x})L_1 + \frac{(\alpha - \bar{x})^2}{2}L_2$$

$$L_0 = \sum_{i=1}^{n} \ln P(x_i, \bar{x})$$

$$L_1 = \sum_{i=1}^{n} \frac{1}{\partial \alpha} \ln P(x_i, \bar{x})$$

$$L_2 = \sum_{i=1}^{n} \frac{1}{\partial \bar{x}} \ln P(x_i, \bar{x})$$

we get maximum likelihood estimate, by calculating

$$L_1 + (\alpha^* - \bar{x})L_2 = 0$$

i.e. $\alpha^* = \bar{x} - \frac{L_1}{L_2}$

\( (D4.4.4) \)

The error in this is $O\left\{ (\alpha^* - \bar{x})^2 L_2 / L_2 \right\}$

which is $O\left( \frac{1}{n} \right)$ as we will soon show that $(\alpha^* - \bar{x})^2$ is $O\left( \frac{1}{n} \right)$.
Now \( L_1 \) and \( L_2 \) are random variables and (D4.4) defines \( \alpha^+ \) as a random variable. We can calculate:

\[
\langle \alpha^+ \rangle = \alpha_0 - \left( \frac{L_1}{L_2} \right)
\]

\[
= \alpha_0 - \left( \frac{\partial}{\partial \alpha} \ln P(x, \alpha) \right) \bigg/ \frac{\partial^2}{\partial \alpha^2} \ln P(x, \alpha)
\]

Use theorem on F/G, proved at end of D2: note this is only valid to \( O(1/n) \). Now the numerator is

\[
\left( \frac{\partial}{\partial \alpha} \ln P(x, \alpha) \right) = \frac{\partial}{\partial \alpha} \int dx \, P(x, \alpha, \alpha) \frac{\partial}{\partial \alpha} \ln P(x, \alpha)
\]

\[
= \int dx \, \frac{\partial}{\partial \alpha} P(x, \alpha) = \frac{\partial}{\partial \alpha} \left[ \int dx \, P(x, \alpha) \right]
\]

\[
= \frac{\partial}{\partial \alpha} \cdot 1 \quad \text{as } P \text{ is} \quad \text{normalized.}
\]

\[
= 0. \quad \text{This proves part (a) that}
\]

\[
\langle \alpha^+ \rangle = \alpha_0 \quad \text{with corrections from both}
\]

\( L_1/L_2 \) and \( L_1/L_2 \) are quite 0 terms which are both \( 1/n \). These corrections
have been calculated by Yellin.

Now to prove part (b): first consider "physics" method. This takes standard deviation of \( \alpha^* \) as \( \sqrt{-\frac{1}{L_2}} \) or rather this evaluated at \( \alpha = \alpha^* \) not \( \alpha = \infty \) but these are same asymptotically.

\[
\sigma^2_{\text{physics}} = -\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\partial^2}{\partial \alpha^2} \ln P(x_i, x_0) \right]
\]

\[
< \sigma^2_{\text{physics}} > = -\frac{1}{n} < \frac{\partial^2}{\partial \alpha^2} \ln P(x_0, x_0) >
\]

where \( < \frac{\partial}{\partial \alpha} \ln P(x, x_0) > = \int dx \ P \left\{ -\frac{2}{p^2} \left( \frac{\partial^2}{\partial \alpha} + \frac{\partial^2}{\partial \alpha^2} \right) \right\} \)

Now second term is just \( \frac{2}{\alpha^2} \int P \) and is zero again because of normalization condition. So

\[
< \sigma^2_{\text{physics}} > = \frac{1}{n} \int dx \left( \frac{\partial^2 P}{\partial \alpha^2} \right)^2 / p \]

\[(D4.5)\]

we can calculate the correct statistical standard deviation in \( \alpha^* \) by taking
(D4.4) and using our $F/g$ theorem with $f_i = \frac{\partial}{\partial a_i} \ln P(x_i, w), \ g_i = \frac{\partial^2}{\partial a_i^2} \ln P(x_i, x)$

Generally

$$\sigma^2 = \frac{1}{n} < \left( \frac{f}{g} \right) - g \left( \frac{f^2}{g^2} \right) >$$

but $<f> = 0$ - proved above, so:

$$\sigma^2 = \frac{1}{n} \frac{<f^2>}{<g^2>}.$$ It is easy to see that $<f^2> = \int \frac{(\partial P/\partial a)^2}{P} p dx$ and this $= <g>$. So $\sigma^2$ statistical is just $1/n<g>$ which is exactly equal to $\sigma^2_{\text{physics}}$ from (D4.5). The expectation value is no sweat: from central limit theorem $\sigma^2_{\text{physics}}$ and $\sigma^2_{\text{physics}}$ differ only by relative error $O(1/n) \to 0$ as $n \to \infty$. 
the simplest problem i.e. $f_i$ are linear in theoretical parameters. Berington should be read before tackling a real problem.

(ii) Note that if $x_i$ are calculated from the same population e.g.

$$x_1 = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$x_2 = \frac{1}{n} \sum_{i=1}^{n} y_i^2$$

Then they are not independent but we can use the theorem proved at the end of Section 3.2 to show that $x_i$ will (for large $m$) be Gaussian but there is a correlation matrix $\Sigma$ i.e.

$$L \propto \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

as discussed in Solmzy.

By diagonalizing $\Sigma$ we can (in principle) reduce this to
the usual $x^2$ method with $x_i$ replaced by eigenvectors of $M$.

This idea is important in the method of moments when one calculates several moments from the same distribution of events. In fact, one usually drops (forgets) off-diagonal terms in $M$ and forms $x^2$ from non-independent reaction variables. This is done in high energy physics when fitting the moments $g_00, g_{11}, g_{20}$ of a 1-particle decay. Of course, in this case, the off-diagonal elements are never given by the experimentalist recording his data; so one is forced to set them zero.
One should also mention that there are two versions of $\chi^2$ depending on whether $\sigma_i$ taken from "experimental" or estimated theoretically. In the latter case, $I_2$ will be a function of the parameters to be varied. For example, suppose $x_i$ is number of events in some counting experiment. Let

$$\sigma_i (\text{experimental}) = \sqrt{x_i} \quad \text{but} \quad \sigma_i (\text{theory}) = \sqrt{E_i}$$

Now $x_i$ is only Gaussian for large $x_i$ but then $\sqrt{x_i}$ and $\sqrt{E_i}$ are essentially the same. If $x_i = 0$, then theoretical $\sigma$ is to be preferred. Although $\chi^2$ method itself is in difficulties as $\chi^2$ not Gaussian, best to go back to maximum likelihood and put in true Poisson distribution for $x_i$. 


Counter Physics

For obvious reasons, I am forced to take my examples from high energy physics.

Take M happy smiling counters all arranged in a row. Suppose they count random events that fall in their province. Then each of them has a parameter \( \lambda_i \) associated with them such that distribution of number \( n_i \) of events in counter \( i \) is Poisson with parameter \( \lambda_i t \) (\( t \) is time). This follows from our theory of Birth processes.

In HEP, \( \lambda_i \propto \int_{BW} dt \cdot \frac{d\sigma}{dt} \) where bin is t-lute covered by counter.
This is a straightforward application of likelihood:

\[ p_i = e^{-\lambda_i t} \left( \lambda_i t \right)^{n_i} / n_i! \quad (D4/3.2) \]

\( n_i \) observed, \( \lambda_i \) theory; \( t \) may be known or not.

\[ \Lambda = \prod_{i=1}^{m} p_i \quad (D4/3.2) \]

or \( \ln \Lambda = - \sum_{i=1}^{n} \ln n_i! - t \sum_{i=2}^{m} \lambda_i + \sum_{i=3}^{n} n_i \ln (\lambda_i t) \)

(ii) Unnormalized Theory Experiment

Suppose \( t \) - which includes beam intensity etc. - is unknown. Then we use \( (D4/3.2) \) to determine \( \Theta \) (regard \( t \) as a theoretical parameter)

\[ \frac{d(\ln \Lambda)}{dt} = 0. \]

or \( t = \sum_{i=1}^{m} n_i / \sum_{i=2}^{m} \lambda_i = N / \Lambda \)

where \( N = \text{total number of events} \)

\( \Lambda = \text{Total Poisson parameter summed over all counts} \)
Now if we forget $t$, we get
\[
\ln L = -\frac{3}{2} \ln n_i! - N + N \ln N + \sum_{i=1}^{m} n_i \ln \left( \frac{\lambda_i}{\lambda} \right)
\]
or dropping terms independent of $\lambda_i$:
\[
\ln L = \sum_{i=1}^{m} n_i \ln \left( \frac{\lambda_i}{\lambda} \right)
\]
(14/3.3)
\[
= \sum_{j=1}^{N} \ln \tilde{p}(C_j)
\]
where $C_j$ is counter where event added and $\tilde{p}(k) = \frac{\lambda_k}{\lambda}$ is probability of landing in $k$'th counter.
Note $\tilde{p}$ is normalized $\sum_{k=1}^{m} \tilde{p}(k) = 1$.
This is discrete maximum likelihood method: we have $N$ observations:
each has only $m$ possible outcomes
and the $L = \prod_{j=1}^{N} \tilde{p}(C_j)$ is product of probabilities - normalized properly.
If $x_i$ are functions of some theoretical parameters, we find them by maximizing (0.4/3.3). If all values of $p_i = x_i/N$ are free parameters, we maximize (0.4/3.3) subject to the constraint $\sum p_i = 1$; this gives us

$$p_i = n_i/N$$

(iii) Normalized Experiment

Now we suppose it is no longer an unknown theoretical parameter, but a known number.

$$\ln L = \text{const} + \left\{ \frac{N \ln (kN) - kN}{k} \right\} + \sum_{i=1}^{3} n_i \ln \left( \frac{x_i}{N} \right)$$

(b) is as before. (a) can be regarded as $\ln p_N$ where $p_N$ is probability of getting $N$ (which should now be...
regarded as an experimentally observed total counts. \( P_N = e^{-\bar{N}} (\bar{N})^N \) which is up to a constant, usual Poisson. This requires some care to apply previous theorems as probabilities must be normalized in proofs in 34.1. However, theorems do hold.

NotePearl calls this extended maximum likelihood method.

\textbf{(iv) Gaussian Limit}

\[ p_i = \exp \left\{ \frac{n_i - \bar{x}_i)^2}{2\bar{x}_i} \right\} \]

we get classical \( \chi^2 \) method. Note if

is theoretical value! (replace \( \bar{x}_i \) in denominator by \( n_i \) to get experimental value). So we only get something different by using maximum likelihood method if some \( n_i \) small and Poisson \( \neq \) Gaussian.
This is like 04/3 except that every event (not every counter) has its own probability $p_i$. (Given events labelled by continuous parameter(s), $p_i$ is a continuous probability density. So far, we haven't discussed how to apply maximum likelihood method to counting problem with continuous probability distribution. (Lifetime example was continuous but not a counting experiment). What we do is divide range of continuous parameter(s) \((0 < \theta < 2\pi)\) in \((1 + \rho \cos \theta) / 2\pi\) polarization distribution into $M$ small bins. $M$ should be so large that (a) probability of 2 events in a given experiment falling in same
lin is zero. (b) Theoretical pi; constant over lin.

Then we have m observations—each with one of two results:

- Success -1 event in lin
- Failure 0 events in lin.

Let N be total number of events. 

Then likelihood is

\[ L = \prod_{\text{successes}} p(\text{success}) \prod_{\text{failures}} p(\text{failure}) \]

\[ = \prod_{i=1}^{N} \exp(-pi \cdot dx_i) (pi \cdot dx_i) \]

\( \prod_{j=1}^{m-N} \exp(-pj \cdot dz_j) \)

Writing out L in terms of Poisson probabilities for 0 or 1 events.

Now m > N and m events dxj cover space. Thus

\[ L = \exp \left[ - \int p(x) dx \cdot t \right] \prod_{i=1}^{N} pi \cdot dx_i \cdot t. \]
Now, fortunately, arbitrary \( dx_i, \)
\( i = 1 \ldots N \) are independent of theoretical parameters and so can be dropped.

We (at further cost to the significance of normalization of \( \mathbf{L} \)).

So our fundamental formula for \( \mathbf{L} \) in a continuous distribution is

\[
\mathbf{L} = \exp \left[ -\int p(x) dx \right] \prod_{i=1}^{N} p(x_i, \alpha) \tag{ii}
\]

Note that, just as in D4/3 -
if \( \tau \) is unknown, we can find \( \tau \) from:

\[
\frac{\partial \mathbf{L}}{\partial \tau} = 0 \quad \text{which gives}
\]

\[
\tau = \sqrt{N / \int p(x) dx}
\]

when \( \alpha \), \( \mathbf{x} \), remaining parameters,

\[
L(\alpha) \propto \prod_{i=1}^{N} \left\{ \frac{p(x_i, \alpha)}{\int p(x, \alpha) dx} \right\}
\]

-the product of normalized probability distributions.
If it is known, the discussion is again similar to this case in D4/3.

(iii) Binned Data

If we have so many events that I can divide them into k groups of n_k events where p constant has same value for each member of group, then:

\[ k = \exp \left\{ -\int_{\mathbb{R}} p \, dx \right\} \prod_{k=1}^{K} \left( pt \right)^{n_k} \]

\[ \alpha = \exp \left\{ -\sum_{k=1}^{K} \lambda_k t \right\} \prod_{k=1}^{K} \left( \lambda_k t \right)^{n_k} (D4/3) \]

\[ \lambda_k = \int p \, d\sigma \text{ lin for k}\text{'th group} \]

and we return exactly to formulae in D4/3. Further if n_k large for each k, our Poisson distributions become Gaussian and we can use \( \chi^2 \) method to analyse data.
This is the only case where $X^2$ correct in sense of giving best answer. However even if $N$ not so large that $p$ constant in each group, one can still bin data and throw away exact values of $x_i$ for each event. Just record what bin it's in. Obviously if $p$ not constant over each bin, this throws away information. However (D4/14.1) is now exact for $x$ and one chooses bin so that $N$ large and $X^2$ applicable. This is used in practice because $X^2$ is so much easier and cheaper to use. Often one can bin data and minimize $X^2$; this gives values for theoretical
parameters which can be used as initial values for expensive full likelihood fit which will then need only a few iterations. One can carry this back further and use a super low grade method (e.g. method of moments - see 54/5) to initialized parameters before $\chi^2$ fit.

This is discussed by Solmity ("histogram method") and Eadie et al.
D4/5 Method of Moments

This is a simple and in general, non-optimal method. Even here - for > 1 parameter - there are two versions (a) ignore correlations (used in practice but wrong) and (b) include correlations (best of a bad lot). Of course simplicity can lead to huge savings in computer time and so it should not be sneered at.

The basic idea of method of moments is simple: consider a set of observations $x_i$, then central limit theorem implies

$$\frac{1}{N} \sum_{i=1}^{N} f(x_i) \rightarrow \mu$$

where $\mu = \int p(x,u) f(x) \, dx$. 
Here $p$ is $p(x)$ where $x$ is some unknown parameter. So putting
$$p(x) = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$
gives one equation for $x$ with solution
$$x = x^*_n; \quad x^*_n \text{ is moment estimate of } x.$$ Note:

(a) One gets a different moment method for each choice of $f$: from
the m.l. theorem, all of these are asymptotically worse or perhaps
just not better than maximum
likelihood method. One must choose $f$ which is simple and has small expected error.

(b) If we have to estimate $n$
th theoretical parameters $a$. Then we
just get $n$ equations by taking
$n$ moments (i.e. $n$ different values of $f$).
(c) Error in \( \mu \) is

theoretically \( \frac{1}{\sqrt{N}} \sqrt{\int f^2 p(x, x) \, dx} \)

experimentally \( \frac{1}{\sqrt{N}} \sqrt{\frac{1}{N} \sum_{i=1}^{N} f^2(x_i) - \left( \frac{1}{N} \sum_{i=1}^{N} f(x_i) \right)^2} \)

using \( \mu_0 \) to be true value of mean

and \( \mu = \frac{1}{N} \sum_{i=1}^{N} f(x_i) \) to be estimate.

By central limit theorem, for large \( N \), experimental estimate will give correct error. However - if all \( x_i \) same (always true if \( N=1 \! \! \! / \) - then experimental error \( = 0 \), which is simply wrong. The moral is: use theoretical estimate if you only have a few events.
(ii) Example:

\[ p(x) = \frac{1}{2} (1 + ax) \quad -1 \leq x \leq 1 \]

(a) **moments method**

\[ \langle x^m \rangle = \frac{1}{2} \int_{-1}^{+1} x^{m+1} \, dx \]

\[ = \frac{a}{m+2} \quad : m \text{ odd} \]

So if \( M_m = \text{moments method using } \langle x^m \rangle \)

\[ a_{-m}^m = (m+2) \sum_{i=1}^{N} x_i^m / N \]

The error squared is

\[ \sigma^2_{m} = \frac{(m+2)^2}{2} \left\{ \int_{-1}^{+1} x^{2m} (1 + ax) \, dx - \frac{a^2}{(m+2)^2} \right\} \]

\[ = \frac{1}{2} \left\{ \frac{(m+2)^2}{2m+1} - a^2 \right\} \]

which is a minimum for \( m = 1 \)

when \( \sigma^2_{1} = \frac{1}{N} (3 - a^2) \)
(a) Maximum Likelihood Method

Now \[ \frac{1}{[N \sigma^2]} |_{m.l.m.} = \int_{-1}^{1} dx \frac{1}{\sigma^2} (3p^2) \]

from our general theory

\[ = \frac{1}{2} \int_{-1}^{1} dx \frac{x^2}{(1+ax)} \]

\[ = \frac{1}{2a^2} \left\{ \ln \left[ \frac{1+a}{1-a} \right] - 2a \right\} \]

\[ = \frac{1}{3} + \frac{a^3}{5} + \frac{a^7}{7} + \ldots \]

\[ = \frac{1}{3} \quad a = 0 \quad \text{which is } \frac{1}{\sigma^2} |_{1} \text{ value} \]

\[ = \infty \quad a = \pm 1. \]

So \[ \sigma^2 |_{m.l.m.} = \sigma^2 |_{1} \quad a = 0 \]

\[ < \sigma^2 |_{1} \quad a \neq 0. \]

(b) Several Parameters

As we have already mentioned, the case of several parameters is treated by taking several moments. Now our moment errors are described by a full second moment matrix for off-diagonal items.
come from \( \int (f(x) - \langle f \rangle)(g(x) - \langle g \rangle)p(x)\,dx \)

which is non-zero in general.

It is conventional to ignore correlations:

the expansion in terms of \( Y_m^k(\theta, \phi) \) of \( d\sigma/d\Omega \) or an angular decay distribution

is typical. The expansion functions \( \{Y_m^k\} \) are orthonormal, so

\[
W(\theta, \phi) = \sum_{k,m} a_{lm} Y_{lm}
\]

\[
a_{lm} \propto \langle Y_{lm}^* \rangle
\]

but still \( a_{lm} \) have correlated errors

as \( \langle Y_{lm} Y_{lm'}^\dagger \rangle \neq \langle Y_{lm} \rangle \langle Y_{lm'}^\dagger \rangle \)

These correlations in HEP example

of \( d\sigma/d\Omega \) express fact that

\[ [1] \]

\( \Theta \)

backward peak (\( \ll \) forward) comes

from correlated cancellations.
In formula

\[ \langle f \rangle = \frac{\sum_{i=1}^{N} f(x_i)}{N} \]  

I used to worry if you should include counting error (\(\sqrt{N}\)) in \(N\) in error estimate which otherwise only involves the standard deviation of \(f\).

If \(N\) is a prior fixed, clearly there is no \(\sqrt{N}\) error but suppose the events used in (**) are those that fell in a particular subregion (corner) of a bigger sample space (and \(f(x)\) is defined only over this subregion). Then estimate of probability of falling into corner certainly involves \(\sqrt{N}\) error! What about mean of quantity within subregion?
The problem is: what is error in a moment: should it include counting error? The answer is no—one uses identical formulae to care when number of evals prescribed and not subjected to counting error.

Show this in a model problem in 2 variables: \( x \) is variable on which moment taken and \( y \) is binned. Choose:

\[
P(x, y) = p(y) \left[ 1 + \alpha(y)x \right]^{\frac{1}{2}}
\]

Take a \( y \)-bin \( y_1 \leq y \leq y_2 \)

Set \( S(y) = 1 \) if \( y_1 \leq y \leq y_2 \)

\[
= 0 \quad \text{otherwise.}
\]

Assume \( \alpha(y) \) is constant at \( x \) over bin and then method of moments is

\[
\alpha = 3 \frac{\sum_{i=1}^{N} f(x_i, y_i)}{\sum_{i=1}^{N} g(x_i, y_i)} = \frac{3F}{\int g(x, y) \, dy}
\]
where \( f(x, y) = x S(y) \)
\( g(x, y) = S(y) \)

Now we can apply F/G theorem from end of (D2) and it is not hard to show that \( f \) and \( g \) separately have counting errors \( \overline{F_n} \) but they are correlated and correlation cancels counting errors, so that \( x \) has identical error formula to standard moment method. (The estimate of \( p(y) \) will involve counting error of course.)
D5: Minimization

D5/1: Problem

The basic problem is to minimize $f(x, \ldots, x_n)$ - a function of $n$ variables $x_1, \ldots, x_n$. Equivalently we must find the zero of an arbitrary function $f'(x, \ldots, x_n)$ of $n$ variables. There arises in previous sections when we wish either to minimize $x^2$ or maximize the likelihood (equivalently to minimize $-\ln L$).

The methods divide according to:

(i) Can calculate $f$ only

(ii) " " $f, f'$ and $f''$

(iii) is extremely time consuming

(i.e. it takes a lot of function evaluations to converge to desired minimum). In all problems I have faced, it is has been simple to
calculate derivatives and so I will only discuss (ii).

5.2 Derivative Formalism

(i) One great advantage of the \( x^2 \) or maximum likelihood problem is that to calculate the second derivative of the function to be minimized only requires the first derivative of the theoretical functions. We first consider this in the case of \( x^2 \):

\[
X^2 = \sum_{i=1}^{N} \frac{(t_i - e_i)^2}{\sigma_i^2} \quad (5.1)
\]

where \( t_i \) is theory function of \( n \) parameters - we take, wolog, \( n=1 \) and one parameter \( \alpha \) in the following:

\[
\frac{\partial X^2}{\partial \alpha} = \sum_{i=1}^{N} 2 \frac{(t_i - e_i)}{\sigma_i^2} \frac{\partial t_i}{\partial \alpha}
\]

\[
\frac{\partial^2 X^2}{\partial \alpha^2} = S_1 + S_2
\]

\[
S_1 = \sum_{i=1}^{N} \frac{2 (t_i - e_i) \frac{\partial^2 t_i}{\partial \alpha^2}}{\sigma_i^2}
\]
\[ S_2 = 2 \frac{N}{\sum_{i=1}^{N} \left( \frac{\partial t_i}{\partial x} \right)^2} / \sigma^2 \]

Now, \( S_1 \) is \( O(\sqrt{N}) \); thus regard it as a random variable which is a function of the \( N \) random variables \( e_1, \ldots, e_N \). By the central limit theorem, \( S_1 \) has mean \( \mu = \frac{1}{N} \sum_{i=1}^{N} \mu_i \) and standard deviation \( \sigma \)

\[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \]

where \( 2(\xi_i - \xi) \frac{\partial t_i}{\partial x} \) has mean \( \mu_i \) and standard deviation \( \sigma_i \).

Now, if \( t_i \) was evaluated with the theoretical parameters \( \mu_i \) and \( \sigma_i \), \( \mu_i \) is some number. So \( \sigma^2 \) is \( O(1) \) and \( S_1 \) \( O(\sqrt{N}) \) as desired.

Meanwhile, \( S_2 \) is \( O(N) \) as it is the sum of \( N \) numbers. Whence, near solution

\[ \frac{\partial^2 \chi^2}{\partial \alpha^2} = S_2 = 2 \sum_{i=1}^{N} \left( \frac{\partial t_i}{\partial x} \right)^2 / \sigma^2 \]
As claimed, second derivatives of $x^2$ only require evaluation of first derivatives of theory. (Only true when $N$ large.)

(i) Let's do the same thing for maximum likelihood method.

\[-\ln L = -\sum_{i=1}^{N} \ln q_i\]

where $q_i$ are probability of our $L=1 \ldots N$ events. Assume $q_i$ are normalized

\[\int q(x)dx = 1\]. This is crucial

but over subtle modifications in e.g. $\theta^{4/3}$ (iii) "generalized" maximum likelihood method can also be handled.

\[-\frac{d^2 \ln L}{d\alpha^2} = \sum_{i=1}^{N} \left( \frac{q_i}{\bar{q}_i} \right)^2 \frac{\delta q_i}{\delta \alpha}^2\]

\[S_1 = \sum_{i=1}^{N} \frac{1}{q_i} \left( \frac{\delta q_i}{\delta \alpha} \right)^2\]

\[S_2 = \sum_{i=1}^{N} \frac{1}{q_i} \left( \frac{\delta q_i}{\delta \alpha} \right)^2\]
Now $S_{1,2}$ are directly of form suitable to apply central limit theorem: thus they are sums over independent random variables taken from same distribution.

$$<S_1> = N < \frac{1}{2} \frac{\partial^2 q}{\partial \alpha^2} >$$

$$= N \int q \cdot \frac{1}{2} \frac{\partial^2 q}{\partial \alpha^2} d\alpha$$

$$= N \frac{\partial}{\partial \alpha} \left\{ \int q d\alpha \right\}$$

$$= N \frac{\partial}{\partial \alpha} (1) = 0$$

So again $S_1$ only has standard deviation term and is $O(\sqrt{N})$: similarly $S_2$ is $O(N)$ and dominates. It follows that exactly same methods can be applied to $\chi^2$ and maximum likelihood fits: in the future we will only mention the former.
The method

lets first generalize the above to n parameters - getting a vector $F_j$ of first and matrix $M_{jk}$ of second derivatives

$$F_j = \frac{\partial \chi^2}{\partial \alpha_j} = \sum_{i=1}^{N} \frac{(t_i - \alpha_i) \partial t_i}{\sigma_i^2}$$

$$M_{jk} = \frac{1}{2} \sum_{i=2}^{N} \frac{\partial^2 \chi^2}{\partial \alpha_j \partial \alpha_k} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial t_i}{\partial \alpha_j} \frac{\partial t_i}{\partial \alpha_k}$$

Note that - in our approximation - $\chi$ is positive semi-definite - which is also a technical advantage.

Now given any guess $\alpha = \alpha_0$

$$\chi^2(\alpha) = \chi^2(\alpha_0) + \sum_j (\alpha - \alpha_0) F_j$$

$$+ \sum_{jk} (\alpha - \alpha_0) (\alpha - \alpha_0) M_{jk}$$
Adopt a vector notation with $\Delta \alpha_j = \alpha_j - \alpha_{j0}$

$$\chi^2(\alpha) = \chi^2(\alpha_0) + \Delta \alpha^T M \Delta \alpha$$  \hspace{1cm} (D5.3)

minimum condition is

$$\frac{\partial \chi^2(\alpha)}{\partial \alpha_j} = 0 \hspace{1cm} \text{or} \hspace{1cm} M \Delta \alpha = -F$$  \hspace{1cm} (D5.4)

so we have new guess

$$\alpha = \Delta \alpha + \alpha_0 = \alpha_0 - M^{-1} F$$

for the minimum. Note that as $M$ is positive definite in (D5.4), this is indeed a minimum as long as second order Taylor expansion valid.

(ii). Error calculation

when we are at a minimum,

$$F = 0 \hspace{1cm} \text{and} \hspace{1cm} (D5.3) \hspace{1cm} \text{becomes}$$

$$\chi^2(\alpha) = \chi^2(\alpha_0) + \Delta \alpha^T M \Delta \alpha$$
or remembering that the likelihood
\[ \lambda = \exp(-\frac{1}{2} \chi^2) \]

\[ \lambda = \lambda(x_0) \otimes \exp(-\frac{1}{2} \Delta x^T M \Delta x) \]

(05.5)
is expected Gaussian form for \( \lambda \). According to D4, \( M \) is just inverse of error matrix i.e.

\[ \langle (\alpha^* - \alpha) (\alpha^* - \alpha)_j \rangle = M^{-1}_j k \]

where \( \alpha^* = \) estimate from minimizing \( \chi^2 \)

\( \alpha = \) true parameters.

In particular "error" \( \sigma_j \) in \( \alpha_j \) is

\[ \sigma_j = \sqrt{(M^{-1})_j} \]

(05.6)

and the other elements in \( M^{-1} \) are the much forgotten correlation parameters. Note that \( \Delta \alpha_j = \sigma_j \) all other \( \Delta \alpha_j = 0 \), produces a \( \chi^2 \) change of 1.
The Real World

For a real problem, the iterative method described in D5-3(b) always diverges; one must invent various artistic devices to ensure the procedure converges.

The simplest reliable algorithm is called Marquardt's method and this is intuitively easy to understand. Consider $\chi^2$ as a function of several parameters.

\[\begin{align*}
\chi^2 &= f(\alpha_1, \alpha_2, \ldots) \\
&= \sum (y_i - \hat{y}_i)^2 / \sigma_i^2
\end{align*}\]
$\chi^2$ will have a bunch of hills and valleys and we are trying to discover the location of the lowest point in valley. If our assumptions are correct then $\chi^2$ is just a real quadratic function and we have found a step that exactly moves to the minimum.

Unfortunately, this is not usually correct as the step $x_0 \to x_0 + \Delta x$ may be too large to support the assumption, that terms in $\frac{\partial^k (\chi^2)}{\partial x^k}$, $k \geq 3$ are zero, to be correct.
Consider the case when there are two variables.

\[ \chi^2 = \chi_0^2 + \text{linear terms} 
    + A_1 (x_1 - x_1^0)^2 + A_2 (x_2 - x_2^0)^2 
    + B_1 (x_1 - x_1^0)^3 + B_2 (x_1 - x_1^0)^2 (x_2 - x_2^0) + \ldots \]

where we have taken the special case where there is no term proportional to 
\((x_1 - x_1^0)(x_2 - x_2^0)\); such a term only confuses the 
current discussion and we will soon see how to put it back in.

The difficulty (lack of convergence of 
iterative scheme) occurs when

\[ A_1 \gg A_2 \]

so that as a function of \(x_1\), \(\chi^2\) we get

and as a function of \(x_2\), we get
with a valley with steep walls in $\alpha_1$ and gentle walls in $\alpha_2$ direction.

The predicted shift from eqn (05.4), $\Delta \alpha_2 \sim 1/\Lambda_2$ and is typically large compared to $\Delta \alpha_1 \sim 1/\Lambda_1$.

Unfortunately, there is no reason to believe that this large shift is well estimated as assumptions that lead to expansion break down.

Terms like $B_2 (\Delta \alpha)^2 \Delta \alpha_2$ are quite unlikely.

We see that estimate is false and we will typically jump out of valley and find

$$\chi^2 (\alpha_0 + \Delta \alpha) \gg \chi^2 (\alpha_0)$$

The method will diverge!
In the above example, the estimate \( \Delta x \), was valid and not spoiled by \( \Delta x_2 \) misestimate. In a real example, there will be mixing between \( x_1 \) and \( x_2 \) due to nonzero \( \frac{\Delta x}{\Delta x_1, \Delta x_2} \). Both shifts \( \Delta x_1, \Delta x_2 \) will be large and species. There is an intuitively obvious solution of this. Consider the second order expansion term.

\[(x - x_0)^T M (x - x_0)\]

\( M \) is a symmetric positive semi-definite matrix. \( M \) can be diagonalized:

\[M \rightarrow P \Lambda P^T\]

\[D = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
0 & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \lambda_n \\
\end{bmatrix}\]
\[ \lambda_j \text{ are eigenvalues of } M \]
\[ p \text{ are eigenfunctions} \]

Put \( \beta = P \alpha \)

Then \( \kappa^2 = \kappa_0^2 + \text{linear terms} + \lambda_1 (\beta_1 - \beta_01)^2 + \lambda_2 (\beta_2 - \beta_02)^2 + \ldots \)

Now the shifts \( \Delta \beta_j \) can be calculated independently and do not affect each other.
Assume that we order the eigenvalues
\[ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0, \]

Then clearly the first few shifts \( \Delta \beta_j \) corresponding to \( \lambda_1 \ldots \lambda_L \), say are reliable and the remainder \( j = L + 1 \ldots n \) unreliable. Suppose we ignore the predictions for these last and set
\[ \Delta \beta_j = \text{Predicted value} \quad j \leq L \]
\[ = 0 \quad j > L. \]

We need a criterion for choosing \( L \) but subject to this, this method is intuitively reasonable.
In a program that I used for many years successfully, I chose $L$ by the largest value such that

$$\lambda_L > r \lambda_1,$$

and $\lambda_{L+1} < r \lambda_1$.

Where initially $r = 0.1$ and $r$ was increased or decreased based on experience i.e. if the resultant $X^2(\beta_0 + 0\beta)$ was indeed smaller (as predicted) or larger than $X^2(\beta_0)$.

This method is equivalent to replacing $M$ by the regularized version.

$$m = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ 0 & \lambda_L & \cdots \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_1 \end{bmatrix} \rho$$

I note that in using this method, I got to look at values of the $\lambda_j$. It was non-uncommon for

$$\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} = \frac{\lambda_1}{\lambda_L} \approx 10^7 \text{ for } n \approx 30$$

You can ask, why one gets such large ratios. Intuitively, the eigenvectors of $M$ with small eigenvalue correspond to linear...
Combinations of $a_j$ that are badly determined so very small $x$'s say that your theoretical formalism is redundant or alternatively the data inadequate.

For instance in fitting a cross section $d\sigma/d(\cos \theta)$ to \[ \sum_{j=1}^{n} a_j P_{j-1}(\cos \theta) \]

then often the coefficients of the higher order Legendre functions are poorly determined and if you choose $n$ to be "too big", then small eigenvalues will develop. In this case, one can cure the problem by reducing the stop index $n$, but in general no such simple solution exists.

Marquardt's method is a simpler technique but with similar goals, one replaces

$$ M \rightarrow M + \delta I $$

where $I$ is the identity matrix.
So means we replace \( M \) by

\[
M = p^T \begin{bmatrix}
\lambda_1 + \varphi & \lambda_2 + \varphi & \ldots & \lambda_n + \varphi \\
\lambda_1 & \lambda_2 & \ldots & \lambda_n
\end{bmatrix} p
\]

Clearly if \( \lambda_n < \varphi \), then the two formalisms are highly equivalent.

Eigenvalues with eigenvalues \( > \varphi \), are shifted by naive prediction while those with eigenvalue \( < \varphi \), have their shift drastically reduced.

One needs an artistic method to estimate \( \varphi \) which again is varied based on experience. \( \varphi \) is increased when predicted \( \chi^2(\beta_0 + \beta) > \chi^2(\beta) \) and is otherwise decreased or left unchanged.

The natural "scale" for \( \varphi \) is given by

\[
\frac{1}{n} \text{Tr} M = \frac{1}{n} \sum_{j=1}^{n} \lambda_j \text{, the average eigenvalue.}
\]
Marquardt's method is less prone than the diagonalization method but much faster; finding eigenvalues and vectors is time consuming.

I have used both successfully.

Diagonalization method for "large" individual fits with many parameters;

I used Marquardt's method in the analysis of some $10^6$ femilab events where in each event I fitted known shape shapes to measured photon energy measurements in a segmented lead-scintillator sandwich. Here $n=10$ and the large number of fits ($10^6$) required a fast method.
G7 $\chi^2$, Bartlett's S function

When the maximum likelihood method reduces to $\chi^2$, it is relatively easy to judge goodness of fit.

$$\chi^2 = \sum_{i=1}^{n} \frac{(e_i - t_i)^2}{\sigma_i^2}$$

where $t_i$ depend on $m$ theoretical parameters. Then the result of minimizing (1) with $m$ theoretical parameters should be distributed according to $\chi^2$ distribution with $n-m$ degrees of freedom. This function is plotted on p.64 of EDJS. If $k$ is number of degrees of freedom.

$$p(\chi^2) = \frac{(\chi_{k/2}^2)^{k/2-1} \exp(-\chi^2/2)}{2^{k/2} \Gamma(k/2)}$$

$$<\chi^2> = k$$

and standard deviation is $\sqrt{2k}$. 
Consider $\chi^2$ with one degree of freedom

$$\chi^2 = \frac{(x-t)^2}{\sigma^2}$$

where $p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{(x-t)^2}{2\sigma^2}\right)$

$$p(\chi^2) = p(x) \frac{dx}{d\chi^2} = \frac{p(x) \sigma^2}{2(x-t)} \frac{dx}{2 \sqrt{\chi^2}}$$

which is value on (*) using $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

One can derive $\chi^2$ distribution for $k$ degrees of freedom in two ways:

1) Mathews and Walker: direct calculation as above noting that $\sum_{i=1}^{k} x_i^2$ mean, std. dev. where $r$ is radial distance in $k$ dimensional space. Then factors multiplying $\exp \left(-\chi^2/2\right)$ in (*) are just volume element in $k$ dimensions.

2) Eadie et al. Show characteristic function of $\chi^2$ with one degree of freedom is

$$\int_{-\infty}^{+\infty} \exp(it \chi^2) p(\chi^2) d(\chi^2) = \frac{1}{\sqrt{1-2it}}$$

... characteristic function for $\chi^2$ with
k degrees of freedom is \((1-2t)^{-k/2}\) and this can be inverse gamma transformed to give \((*)\).

Introduce the characteristic function \(F(x^2)\)

\[
F(x^2) = \int_0^\infty p(x^2) \, d(x^2) = \frac{1}{2 \Gamma(k/2)} \int_0^\infty (\frac{x^2}{y_2})^{(k/2)-1} e^{-y/2} \, dy
\]

\(F\) has probability distribution \(\omega(F)\)

\[
\omega(F) = \frac{p(x^2)}{d(x^2)/dF} = 1
\]

So \(F\) is uniformly distributed between 0 and 1.

We prefer to define \(C(x^2) = 1 - F(x^2)\), which is called the confidence level

\[
C(x^2) = 1 \quad \text{is} \quad x^2 = 0 \quad \text{"good"}
\]

\[
= 0 \quad \text{is} \quad x^2 = \infty \quad \text{"bad"}
\]

or "low confidence"
Suppose you conduct a series of experiments which gives you a sequence of $X^2$ values

\[ X^2 \mid 1, 2 \ldots \]

An example in high energy physics would come from analysing a set of events each containing particle tracks - each track would have an individual $X^2$. A given experiment would generate many million such $X^2$ values!

If one can do the "emus" right, the resultant distribution in $C(X^2)$ should be flat

\[
\begin{align*}
C(X^2) & \text{ flat} \\
& 0 \leq X^2 \leq 1
\end{align*}
\]

but if you have "background" i.e. events that do not satisfy claimed hypothesis, then one will have an excess of events at large $X^2 \gg 1$ or small $C(X^2)$
$\binom{X^2}{Y}$ will recommend look like

The plot of $\binom{X^2}{Y}$ is a good way of estimating background.

Unfortunately one often gets $\binom{X^2}{Y}$ plots that look like

with no clear "flat" piece. This may be because your "signal" is very small. It could be because you got your errors wrong (?)

E.g. maybe the true $\sigma_i$ is are
all larger than ones used in calculating $\chi^2$. In this case it makes sense to scale $\chi^2$ ($\chi^2_{\text{new}} = c \chi^2_{\text{old}}$) until we get a nice flat distribution for $C \approx 1$ and analyse as on top of p. 19.

If we expand $\chi^2$ near its minimum, with for convenience - just one theoretical parameter -

$$\chi^2 = \chi^2_0 + \alpha(\lambda - \lambda_0)^2$$

where $\lambda_0$ is our estimate of $\lambda$. Then probability distribution of $\lambda$ is just

$$\lambda \sim \exp \left(-c \frac{\chi^2(\lambda)}{2}\right)$$

$$\alpha \sim \exp \left(-\alpha c \frac{(\lambda - \lambda_0)^2}{2}\right)$$
\[ \frac{1}{\sqrt{n}} \text{ is standard deviation of estimate. Returning to } (***) \text{, we find that if we put } \\
\lambda = \lambda_0 \pm \frac{1}{\sqrt{n}} \text{ (i.e. at its one standard deviation value), } \chi^2 \text{ is increased by 1. This } \chi^2 \text{ increase "rule" is } \\
\text{common method of finding "error" in estimate } \lambda \text{ even when the Taylor expansion (***) is not a } \\
good approximation. \\
\text{This can be generalized to } \\
\text{Bartlett's } S \text{ function which is defined as } \\
S(\lambda) = \frac{1}{\sqrt{I(\lambda)}} \frac{\partial \ln I}{\partial \lambda} \\
I(\lambda) \text{ is information } < (\frac{\partial}{\partial \lambda} \ln I)^2 > \]
\[ I(\lambda) = N \int \left( \frac{d}{dx} \ln p(x, \lambda) \right)^2 p(x, \lambda) \, dx \]

\[ \text{can be calculated from theoretical forms.} \]

\[ \text{Number of events.} \]
From \((\star\star)\) we have
\[
\frac{1}{3h} \ln L = -\frac{2a(x - \lambda_0)}{2} \quad (L = \exp(-Z^2/2))
\]
and using our earlier asymptotic result \(a = \sqrt{I(x)}\), we find
\[
S(x) = \sqrt{I(x)} (\lambda_0 - x)
\]
\[
= \sqrt{a} (\lambda_0 - x)
\]
\[
\therefore \quad x = \lambda_0 \pm 1 \text{ Standard deviation is just } S(x) = \pm 1. \text{ In fact, for any } n, S(x) \text{ has mean } 0 \text{ and variance } 1 \text{ and so the rule } S(x) = \pm 1 (\pm \ell \text{ for } \ell \text{ Standard deviations}) \text{ is an appropriate generalization of that to increase } x^2 \text{ by } \ell^2 (\ell).
Hypothesis Testing

We wish to test a specific hypothesis $H_0$ (referred to as null hypothesis) against a set of possible other hypotheses. In goodness of fit analysis, we test $H_0 (\neq$ hypothesis that data described by claimed theory with parameters found from maximizing $L$) against "all" other hypotheses. If we are trying to find type of a particle then $H_0$ might be particle is an electron, $H_1$ that is a pion and these could be only possibilities. We are given
our standard N observations (x₁, ..., xₚ) on which to make our decision. We do this by constructing yet another statistic Y which is some function of x₁, ..., xₙ. We assume that Y has distribution Pₓ(i) if hypothesis i is true (i = 0 or 1 if 2 hypotheses). Y could be vector valued but we will take scalar case here. In the rest of all possible worlds, one can find a Y such that the values taken by it under two hypotheses are disjoint i.e.

\[ \int Pₓ(\theta) Pᵧ(i) dy = 0 \quad (i) \]
Unfortunately, (4) is not always attainable but one should always try to minimize the probability overlap integral. We define $\mathfrak{r}$ as region $\mathfrak{g}$ in $Y$ space where we will deem $Y$-i.e. hypothesis (1) to be true and (0) to be false. For the remaining region of $Y$ space, $R - r$, (R total $Y$ range) we will take (0) to be true, (1) to be false. This scheme is characterized by 2 constants:

$$\alpha = \text{level of significance or size of test} = \int_{r}^{R} P_{Y}^{(0)}(y) \, dy$$

$\alpha$ is chance that (0) will be
declared false when it is in fact true.

\[ \beta = \int_{-\infty}^{\infty} p_y(y) dy \] is the contamination

i.e. chance that \( H_0 \) will be declared
true when it is in fact false. Clearly
it depends on the application.

The balance between efficiency
1-\( \alpha \) and contamination \( \beta \).

Student's t distribution

\( x_i \) are drawn from a
normal distribution of unknown
mean and standard deviation.

\( H^{(0)} \) is \( \text{mean} = \mu^{(0)} \)

\( H^{(1)} \) is \( \text{mean} = \mu^{(1)} \).
Define \[ t = \frac{(\bar{x} - \mu)}{b} \]
where \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]
\[ \sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

Under hypothesis (i), \( t \) follows a so-called Student's t-distribution with \( n-1 \) degrees of freedom. \( (f = n-1) \). The probability density is

\[ f(t) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\sqrt{f\pi} \Gamma\left(\frac{f}{2}\right) \left(1 + \frac{t^2}{f}\right)^{(f+1)/2}} \]

\( f = 1 \) is Cauchy distribution

\( f \to \infty \) it becomes Gaussian with mean 0 and standard deviation 1.
One would say that
0 is true if $t < t_a$
1 is true if $t > t_a$.

The corresponding efficiency / contamination are nicely calculable from cumulative probability functions of the Student's $t$ distribution. As discussed on page 168 of Bryant, (use average $\bar{x}$), this method can be adapted to test if 2 sets of measurements have the same mean (e.g. set 1: Smokers set 2: non Smokers).
Robust Estimation


In the maximum likelihood method we found the method that had the smallest possible standard deviation (asymptotically at least). However as we emphasized in the study of non-parametric estimates, it is quite possible (and indeed probable) that systematic error from incorrect theoretical form of probability is large. Even when we knew main functional dependence (Boltzmann for a resonance, say) there are often "small" contributions to the
probability distribution (e.g. background resonance) which must be estimated as polynomial/phase space mass.

which must be parametrized in an ad hoc fashion. It is hard to evaluate distortion due to effect of an incorrect parametrization of these unknown contributors. There is thus a great deal of interest in the study of so-called robust techniques which are insensitive to errors in our knowledge about the underlying distribution. Statisticians are very interested in such techniques—but (apart from non-parametric methods we have already discussed) there has been little use/experience of them.
The type of in-hand problems that physicists are wont to find. Further the derivation of robust methods is rather intuitive and not as mathematically precise as those of classical statistics.

The first book cited has an unsuitable study of a very simple problem; give a unwarranted sample $\{x^{(i)} \mid i = 1, N\}$ from some distribution, find an estimate of the mean $\mu$ of the underlying probability distribution. The simple natural method is to estimate $\mu$ from the arithmetic mean

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

As long as the underlying distribution has a finite standard deviation this will have an error $\sigma / \sqrt{N}$ which
\( \frac{\hat{\mu}}{2} \) as \( N \to \infty \). Further the method is unbiased for any \( N \)

\[ <\hat{\mu}_{am}^* > = \mu. \]

If the \( x^{(i)} \) are distributed according to either the Gaussian

\[ p(x) \propto \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \]

or exponential form

\[ p(x) \propto \exp(-x/\mu) \]

then we know from maximum likelihood method that the arithmetic mean is the best method. However for other distributions, the arithmetic mean is unlikely to be best. For a trivial example, suppose I measured \( \frac{1}{2} x \) where \( x \) had an exponential distribution. Then the best estimate of the mean of...
\( y \) is obviously
\[ y^{\text{true}} = \left( \frac{1}{N} \sum_{i=1}^{N} (y^{(i)})^2 \right)^{1/2} \]
and not \( y^{\text{true}} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)} \).

The problem of estimating \( \mu \) is hardest for distributions \( p(x) \) that have long tails. Then with a small probability we will generate events from the tail but these will be way off the mean and if included in calculation of mean will distort results in a finite sample. One method that is more robust than the arithmetic mean, is the trimmed mean
\[ \mu^{\text{trim}} = \frac{\sum_{i=pN+1}^{N(1-p)} x^{(i)}}{N(1-2p)} \]
where PN events are left off each end. As long as $p$ is fixed $< 0.5$, for the special case of distributions $P^{\text{trim}}$, symmetric about their mean, the trimmed $p^*$ will also have mean $p$ and error $\frac{1}{\sqrt{N}}$, for fixed $N$. The variance of $P^{\text{trim}}$ will be larger than $P^{*\text{am}}$ for Gaussianly distributed $x^{(i)}$. But as you investigate distributions broader than Gaussians you will find $P^{*\text{trim}}$ becoming smaller than that of $P^{*\text{am}}$. A special case of $P^{*\text{trim}}$ is the median (i.e. middle value of $x^{(i)}$ when they are numerically ordered) gotten with $p = 0.5$.

An interesting class of robust estimators are based on the modification of the maximum likelihood equations.
Thus let \( z \hat{\omega} = \frac{x^{(i)} - \mu}{\sigma} \).

and suppose that \( p(x, \theta) \) is just a function of \( z \). (True for Gaussian case!) The likelihood equation for \( p \) is

\[
\frac{\partial}{\partial \mu} \ln L = 0
\]

or

\[
\sum_{i=1}^{N} \frac{\partial}{\partial \mu} \ln p \left( \frac{x^{(i)} - \mu}{\sigma} \right) = 0
\]

Now, as we are assuming \( p \) to be an even function of \( z \), i.e., symmetric about the mean - we write the above as

\[
\sum_{i=1}^{N} f \left( \frac{x^{(i)} - \mu}{\sigma} \right) = 0 \tag{\#}
\]

where \( f \) is an odd function of its argument.

For the case of Gaussian \( p \), \( f \) is just \( \frac{1}{\sigma^2} (x^{(i)} - \mu) \) and (\#) is

\[
\frac{1}{\sigma^2} \sum_{i=1}^{N} (x^{(i)} - \mu^*) = 0
\]

which gives the arithmetic mean as the maximum likelihood estimate.
of the mean. Now we invent a function \( f \) which is less sensitive to \( x^{(i)} \)'s far from the mean than \( (x^{(i)} - \mu) \).

\[ f(z) \]

\[ \text{Huber's Modification} \]

\[ z_c \sim 1 \text{ to } 2. \]

Many forms for \( f \) have been proposed. The most successful was that due to Andrews with

\[ f(z) = \text{Sm} \left( \frac{z}{2.1} \right) \]

\[ = 0 \quad \quad |z| > 2.1\pi \]

\[ -1 \quad \rightarrow z \]

Note \( z \) is \( \frac{x - \mu}{\sigma} \) where \( \sigma \) supplies the natural scale for the problem. We do not need a very good estimate for \( \sigma \) but
it should be robust! A simple choice is \( \sigma = \text{median} |x^{(i)} - p| \) which is applied iteratively as value of \( p \) is refined.

A large number of tests are described in just reference to this section.

Let \( N = \text{Gaussian Distribution} - \sigma = 1 \)
\( C = \text{Cauchy} \)
\( M = \text{Gaussian \pm uniform random variable between 0 and 1} \).

Here is a sample of their results for 10 observations:

<table>
<thead>
<tr>
<th>Method</th>
<th>( N )</th>
<th>( C )</th>
<th>( M )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>27.31</td>
<td>52.55</td>
</tr>
<tr>
<td>Median</td>
<td>1.37</td>
<td>3.66</td>
<td>7.33</td>
</tr>
<tr>
<td>Andrews</td>
<td>1.08</td>
<td>4.66</td>
<td>9.29</td>
</tr>
</tbody>
</table>

The above table gives variances in units of which I am not certain; but the relative values within a column are independent of this! We see very clearly that the arithmetic mean is hopeless for...
Andrews methods work well in these cases but with the Andrews being better in Gaussian case (or only 4% more than best estimate) but less robust in other examples.

Note that all the methods except the arithmetic mean require all the \( x_i \) being in memory so that we can order them. So the computer complexity is much higher for the robust estimates. The book gives computer programs for these cases all the robust estimators.
MONTE CARLO INTEGRATION

This expands discussion on page 23 (Section (iii) of Section 1.7) of typed written notes.

5/4 Monte Carlo Integration

(i) As we will find out later in the course (when we discuss Simpson's Rule and Gaussian Integration), ALL integration methods can be written as

\[ I = \int_{a}^{b} f(x) \, dx = \sum_{i=0}^{n} f(x_i) \, w_i \]

The \( n+1 \) points \( x_i \) and weights \( w_i \) are determined.
E5/4 Monte Carlo Integration

1. "All" integration methods can be written as

\[ I = \int_a^b g(x)f(x)\,dx = \sum_{i=0}^{n} w_i f(x_i) \]

where the n+1 points x_i and n+1 weights w_i are determined

in sundry ways to optimize various properties as described earlier. If we increase n, the dependence of error is

\[ \text{Error} \propto \frac{1}{n^2} \]

\[ \frac{1}{n^4} \]

\[ \frac{1}{2n} \]

\[ \frac{1}{n^m} \]

\[ \frac{1}{n^r} \]

iterated Trapezoidal

iterated Simpson

iterated m-order Gauss

with iterated groups of k points
Taking \( g(x) = 1 \) for convenience, we write

\[
I = (b-a) \int_{-\infty}^{+\infty} f(x) \cdot p(x) \, dx
\]

\[
p(x) = \frac{1}{(b-a)} \quad \text{for} \quad a \leq x \leq b
\]

\[
= 0 \quad \text{outside} \quad [a, b]
\]
Now consider the Monte Carlo which writes \( I = \langle \tilde{f} \rangle \)

\[
\langle \tilde{f} \rangle = \int_a^b \tilde{f}(x) \, dx / (b-a) \quad \tilde{f} = (b-a) f
\]

\( I \) is mean of \( f(x) \) where \( x \) is uniformly distributed random variable between \( a \) and \( b \).

Defining

\[
\sigma^2 = \int_a^b [\tilde{f}(x) - I]^2 \, dx / (b-a)
\]

we know from central limit theorem that
\[ \frac{1}{n} \sum_{i=1}^{n} f(x_i) \rightarrow I \]

with error \( o(\sqrt{n}) \).

(Eq. 5.11) is of the same form as our previous formulae: all the \( w_i \) are equal, while the positions \( x_i \) are random. However, it seems a bad idea as error is much slower than even iterated trapezoidal.

But wait! Consider an integral in \( k \) dimensions - for other methods, we must make a grid in each direction with \( n^{1/k} \) points for each coordinate. Thus error for Trapezoidal is \( \propto \frac{1}{n^{2/k}} \) which has a slower \( n \) dependence as \( k \) increases. However, Monte-Carlo method still keeps its \( \frac{1}{\sqrt{n}} \) behaviour and for \( k > 5 \), converges faster than Trapezoidal method. The other methods appear better than Monte-Carlo until larger \( k \). However, in practice, this is not so because there the implicit
condition $n^{1/k} \gg 1$ before the error
formulas even apply. For instance
$n^{1/k} = 10$, $k = 5$ gives ridiculous values of
$n$.

We cannot study it here, but this
example shows that the problem of
multidimensional integrals is not a
trivial generalization of that in one dimensions.

For one general problem of best
method in $k$ dimensions: the answer is
surely not either Monte-Carlo or
Gaussian for each co-ordinate. There is
a feasible account in Isaacson and Keller
(chapter 7, p. 352)

(iii) Complicated Domains

One advantage of the Monte Carlo
method is that it can very easily take
account of complicated integration domains.

Suppose we wished to
calculate the integral
over the inside of the circle
\[ I = \int_{x^2+y^2 \leq 1} g(x,y) \, dx \, dy. \]

One way of doing this is to put
\[ g'(x,y) = g(x,y) \quad x^2+y^2 \leq 1 \]
\[ = 0 \quad x^2+y^2 > 1 \]

and calculate
\[ I = \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} g'(x,y) \, dy \, dx \]

which is an elementary Monte Carlo integral with \( x \) and \( y \) generated independently between \(-1\) and 1.

Actually, the finite force method (generate \( x \) uniformly between \(-1\) and 1, \( y \) uniformly between \(-\sqrt{1-x^2}\) and \(\sqrt{1-x^2}\)) is more efficient - in the language of (55.11) the second method has a smaller variance.

However, this is an important technique for it is often impossible to find ranges analytically while
It is easy to decide if a given \((x, y)\) lie within range of integration.

(iii) Importance Sampling

Of course it is obvious from (E5.11) that Monte Carlo integration is better, the smaller is \(\sigma\). Using definition of \(\sigma^2\), (or remembering its name "variance") it follows that method is best when \(f(x)\) is slowly varying over domain. [for instance if \(f(x)\) is constant, \(\sigma = 0\) and method is exact]. Importance Sampling ("sampling where integrand important") is a technique for reducing variation of \(f(x)\).

Thus if \(I = \int f(x) \, dx\)

Change variables \(y = q \cdot f(x)\)

\[dy = q f'(x) \, dx\]

Put \(f \frac{f(x)}{q f'(x)} = \frac{F(y)}{F'(y)}\)

Then \(I = \int F(y) \, dy\).
We now apply Monte Carlo method generating y uniformly.
we try to choose $q'(x)$ so $\frac{f(x)}{q'(x)}$
has a smaller standard deviation
than $f(x)$.
Best choice
$$q'(x) = f(x)$$
when
1) $F(y)$ is constant with zero standard deviation
2) probability of getting a particular
x value is $x q'(x) = f(x)$
i.e. one "chooses" x values to
congregate where function $f$ is large
This is of course not realistic as to find $q'(x) = f(x)$ requires

$$q(x) = \int_a^x f(x) \, dx$$

i.e. solving integral.

In general one chooses $q'(x)$ so that

a) $q'(x)$ "approximately" $f(x)$

b) $\int q'(x)$ can be performed.
Example

\[ f(x) = e^{-x} \otimes \text{a mess} \]
where \text{mess} slowly varying

choose \[ q'(x) = e^{-x} \]
\[ q(x) = 1 - e^{-x} \]

Then \[ F(y) = "\text{a mess}" \]
One can apply Monte Carlo methods to sums as well as continuous integrals.

e.g. Suppose we have

\[ \int dx_0 dx_1 dx_2 \ldots \sum_{l=1}^{2} f_l(x_0, x_1, x_2 \ldots) \]

Then generate

- \( x_a \) with random number
- \( x_b \)
- \( x_c \)

Generate another random number

- \( 0 \leq \xi \leq 1 \) uniformly

if \( \xi < 0.5 \) then \( l = 1 \)
else \( l = 2 \)

and obvious modifications for other discrete summations.
Accept/Reject Method

We wish to generate $x$ according to $p(x)$, $x_1 \leq x \leq x_2$

$$p_{\text{max}} = \max_x p(x)$$

Let $w_{\text{max}} \geq p_{\text{max}}$ (any number)

Generate $x$ uniformly in $[x_1, x_2]$ in $[0, 1]$

if $r > \frac{p(x)}{w_{\text{max}}}$ ignore $x$

$\leq \frac{p(x)}{w_{\text{max}}}$ accept $x$
Metropolis Method (Journal of Chemical Physics 21, 1087 (53))

Given a point \( x_i \) — initially chosen at random, choose a new point \( x_j \) as follows: Choose \( x_j \) in the neighborhood of \( x_i \) uniformly in the \( \int p(x) \, dx \)

if \( p(x_{\text{new}}) > p(x_i) \), accept it, and set \( x_{i+1} = x_j \)

if \( p(x_i) < p(x_j) \), generate a random number \( \Gamma \) uniformly in \([0,1]\)

if \( \Gamma < p(x_j)/p(x_i) \) set \( x_{i+1} = x_j \)

if \( \Gamma \geq p(x_j)/p(x_i) \) set \( x_{i+1} = x_i \)

Eventually this produces a set of points distributed "according to \( p(x) \)."
1) The neighborhoods $\Delta x = x_i \leq x_{i+1}$ eventually cover the whole space.

2) Take any two points $x_i, x_j$. Suppose $p(x_i) > p(x_j)$.

- $\Pr (i \rightarrow j) = \frac{p(x_j)}{p(x_i)}$
- $\Pr (j \rightarrow i) = 1$

\[ \text{i.e. } \frac{\Pr (i \rightarrow j)}{\Pr (j \rightarrow i)} = \frac{p(x_j)}{p(x_i)} \]

In "equilibrium":

\[ N_i \cdot \frac{\Pr (i \rightarrow j)}{\Pr (j \rightarrow i)} = N_j \cdot \Pr (j \rightarrow i) \]

So we have generated $X_i$ according to $p(x_i)$, i.e. exact importance sampling according...
One is evaluating integrals of the form

\[ I(\Omega) = \int \theta(x) p(x) dx \]

Here vector \( x \) is in parameter space of experimental events e.g. in high energy physics \( x \) labels momenta of final state particles.

- \( p(x) \) is probability that a given \( x \) is formed e.g. in high energy physics, \( p(x) \) is square of modulus of production amplitude. \( \theta(x) \) is a particular experimental observable e.g. energy into a calorimeter of a particular geometry.

Typically \( p(x) \) is fixed by the theory but one wishes to calculate \( I \) for several choices of \( \theta \) as one explores different facets of the
exponential. Thus although the result of any given I depends on $\sigma$ of product $\Theta p$, to get good results for many different $\Theta$'s suggests that we minimize the variance of $p(x)$. There are three methods.

1) "Importance Sampling" i.e. change variables to $y$ where the Jacobian $\frac{d(x)}{d(y)} = j(x)$

$$dx = j(x) \, dy$$

and we try to choose $j(x)$ so that variance of $j(x)p(x)$ is minimized - the best choice is of course $j(x) = (p(x))^{-1}$.

2) Metropolis Method.

3) Uniformize "weights".

Suppose

$$W = \max_x [p(x)]$$
Then having generated $x$ uniformly, choose a random number $r$ between $0$ and $1$

$$\begin{align*}
&\text{if } r < \frac{p(x)}{W} \text{ keep } x \\
&\text{if } r > \frac{p(x)}{W} \text{ reject it.}
\end{align*}$$

As events are accepted with probability $p(x)$, they are "distributed according to $p(x)$".

$$I(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \Theta(x_i) : x_i \text{ chosen as alive.}$$

This method is particularly good if it is very hard to calculate $\Theta(x)$. This happens say if response of particular part of apparatus is hard to calculate. (It might involve tracking protons through a magnetic field etc). Then $it$ is "waste" to do a lot of work for events that
are only going to get a small weight
$p(x)$ in the final analysis.
In hard problems, $W$ can often
not be precalculated and must be
found from a preliminary examination
of the integrand. Also the acceptance
rate may be very poor e.g. a fraction
$<p(x)> / W$ is kept and this can be
small (.01 is quite typical in problem
I have worked on).
Lattice Monte Carlo

A typical "theoretical" calculation in condensed matter or high energy physics involves integrals of form

$$I = \int \exp\left[-\frac{H}{kT}\right] \varnothing \, ds_1 \ldots \, ds_N$$

where $s_i$ is spin (gluon field) at site $i$. Typically the sites are in a regular lattice in 2, 3 or 4 dimensions. This integral would involve some $10^4 \rightarrow 10^6$ dimensions and so must be evaluated by Monte Carlo.

The Hamiltonian $H$ is proportional to size of system and so the weighting function $\exp\left[-\frac{H}{kT}\right]$ is very small and one must be very careful to choose one integration points appropriately.

Assume you have a particular choice

$s_1^{(0)} \ldots s_m^{(0)}$ with
\[ H^{(0)} = H(S_i^{(0)}, \ldots, S_n^{(0)}) \]

Then we try to move a small way from \( S_i^{(0)} \) by just changing one of the spins:

\[ S_k^{(0)} \rightarrow S_k^{(1)} \]

\[ S_i^{(0)}, \quad i \neq k, \quad \text{unchanged} \]

Typically, this produces a change \( \Delta H \approx kT \) as \( kT \) is energy per site (upto numerical constants) — remember classical result of energy of \( \frac{1}{2} kT \) per degree of freedom.

One usually uses one of two methods to change \( S_k \):

(a) Metropolis — generate \( S_k^{(1)} \) uniformly about \( S_k^{(0)} \) and accept or reject change based on \( r = \exp \left[ - \frac{[H(\ldots, S_k^{(1)}, \ldots)] - H(\ldots, S_k^{(0)}, \ldots)}{kT} \right] \)

if \( r > 1 \) (energy decreased), accept \( S_k^{(1)} \)

if \( r < 1 \) (energy increased), accept \( S_k^{(1)} \) with probability \( r \).
(b) He or Bath - Sometimes one can use "exact" method to generate a variable according to a given probability distribution i.e. if distribution function

\[ F_k(x) = \int_{-\infty}^{\infty} \exp \left( -H(\ldots, S_k, \ldots)/kT \right) \]

can be calculated and inverted. Here we view \( H \) as a function of \( S_k \) with other spins fixed.

In each case (a) or (b), spins are generated according to distribution \( \exp [-H/kT] \) and so values of observables are calculated as

\[ \langle O \rangle = \frac{1}{N} \sum_{l=1}^{N} O(S^{(l)}_1, S^{(l)}_m) \]

If we had not been careful about generating the spin values
But say chosen them uniformly, then

\[
\langle \Omega \rangle = \sum_{l=1}^{N} \exp \left[ -H [ s^{(l)} ] / kT \right] \Theta ( s^{(l)} )
\]

\[
\sum_{l=1}^{N} \exp \left[ -H [ s^{(l)} ] / kT \right]
\]

This is correct but numerically hopeless. The wide variation in exponential (remember $H$ proportional to $S_{12}$) would give a huge standard deviation in $\langle \Omega \rangle$. 
Generation of Random Numbers

See
Knuth "The Art of Computer Programming"
Volume 2, "Semi Numerical Algorithms"

Knuth is a reasonable pedagogical description although it is a little out of date - Chin describes newer algorithms which are imported in some cases. Press implements algorithms at about the level of Knuth.

One needs specialized hardware to generate true random numbers e.g.
\[ \theta \]

the angle \( \theta \) is random

for successive decays of a bunch of 1's. There are a lot of difficulties here; not
only must you measure $\Theta$ in an unbiased fashion (impossible), to find a random number with 32 bits of precision would require several separate $\Theta$ values - maybe each gives 4 bits of information and so 8 values are needed!

Rather are usually a deterministic method which gives "essentially" random numbers. Here are called pseudorandom.

Most "standard" library random number generators use the so-called linear congruential generators which generate a sequence of integers $I_j$ by

$$I_{j+1} = [a I_j + c] \mod (m)$$

Here $m$ is called the modulus,

$a$ " " multiplier

and $c$ " " increment

This is deterministic - given any seed $I_0$, one generates an identical sequence $I_0, I_1, I_2, \ldots$

There are most $m$ possible random numbers but there may be blank, and could generate a cycle if
\[ I_{j_1} = I_{j_2} \quad \text{if} \quad j_1 > j_2 \]

Then \( I_{j_1+1} = I_{j_2+1} \) etc.

also \( I_{j_1-1} = I_{j_2-1} \)
and in particular \( I_{j_1-j_2} = I_0 \).

If the cycle has a length \( M \), then imagine our calculation of \( \Pi \)

\[ \Pi = \int_{-1}^{+1} \int_{-1}^{+1} dxdy \Theta \left[ 1 - x^2 - y^2 \right] \]

Then random numbers after the \( M \)th
are no longer independent and so our use of central limit theorem
breaks down.

Our error is \( \max \left[ \frac{1}{\sqrt{N}}, \frac{1}{\sqrt{M}} \right] \)

\[ \text{number of events} \]

\[ \text{cycle length} \]
The references give various formal criteria to guide the choice of \( a, c \) and \( m \). Unfortunately, I don't see any of these references as being very good as they don't test and compare some key choices.

In my group we have used

\[
m = 2^{31}
\]

\[
a = 1103515245
\]

\[
c = 12345
\]

with effective C code

\[
\text{newran = } [a \times \text{oldran} \mod m] \mod \text{MASK}
\]

where \( \text{MASK} \) has lower 30 bits = 1

all higher bits = 0.

we get a floating point number between 0 and 1 by:

\[
\text{float ran = } \text{newran} / 2147483648.0
\]

This is very simple to code especially in C which allows you to use machine capabilities effectively.
The Shift register methods are not discussed (correctly) in Knuth and may well be better as they are as easy to calculate and have longer period.

Chin has implemented a particularly simple method.

\[ I_j = I_{j-q} \oplus I_{j-p} \]

where \( \oplus \) = Exclusive OR

\[
\begin{align*}
1+1 &= 0 \\
1+0 &= 0 + 1 = 1 \\
0+0 &= 0
\end{align*}
\]

This \( \oplus \) operates independently on each bit of the numbers \( I_{j-q}, I_{j-p} \).

The period of this could be as large as \( 2^p - 1 \) which is big for Chin's choice \( p = 250, q = 147 \).

Initialization

The Shift register method is unpopular because it is newer and because it is harder to initialize. However, Chin describes a reasonable way using a congruential generator.

Typically a random number package...
includes a call to a routine with a name like: ranset

A call ranset (seed)

Sets $I_0 = \text{Seed}$ for the congruential method

If you want a truly random start, a standard "dodge" is to set the seed
to the linear image (remember $I_0$ is just a (32 bit) integer) of the system clock.

One may wish to redo a calculation with the same initial seed so use

clock (seed)

print (linear, seed)

ranset (seed)

Correlations:

Often purely chosen random number generators exhibit correlations
between neighbouring random numbers

This is well illustrated by tests by Chin of integer generator
Supplied by Bostand with Turbo Pascal

v. 3-0
Fig. 2. More than 1,000,000 random numbers generated by Turbo-Pascal v3.0 and plotted as \((x, y)\) pairs.

Other generators are better!

Numerical Recipes has a program that removes any such problems.

**Essentially set up a list** \(R(l)\) **of** \(L\) **random numbers.**

**Then choose** \(j\) \(1 \leq j \leq L\)

**Select desired random number** = \(R(j)\)

**Replace** \(R(j)\) **by a new random number.**

**Perhaps**

These correlations are particularly important in accurate calculations of condensed matter phase transitions as you are measuring correlations between local spins and so these must have uncorrelated random generators.
Gaussian Random Numbers:

\[ g = \frac{\sum_{i=1}^{N} u_i - N/2}{\sigma \sqrt{N}} \]

where \( \sigma^2 = \frac{1}{2} \) and \( u_i \) are uniformly distributed between 0 and 1, is Gaussianly distributed for large \( N \). This for \( N \approx 10 \) is sometimes used in computers.

More convenient is often:

\[ g_1 = (-2 \ln U_1)^{1/2} \cos 2\pi U_2 \]
\[ g_2 = (-2 \ln U_1)^{1/2} \sin 2\pi U_2 \]

where \( u_i \) are uniform in \([0,1]\)

and then \( g_i \) are independent random numbers Gaussianly distributed with zero mean and unit variance. You can prove this by taking polar co-ordinates in

\[ e^{g_1^2 + g_2^2/2} \]
\[ dg_1 \, dg_2. \]
\[ g_1 = r \cos \theta \]
\[ g_2 = r \sin \theta \]

\[
\exp \left( -\frac{1}{2} \left[ g_1^2 + g_2^2 \right] \right) \, dg_1 \, dg_2
\]
\[ = \exp \left( -\frac{1}{2} r^2 \right) \, dr \, d\theta . \]

\[ = \exp \left( -\frac{1}{2} r^2 \right) \, d \left( \frac{1}{2} r^2 \right) \, d\theta . \]

\[ = d \left( \exp \left( -\frac{1}{2} r^2 \right) \right) \, d\theta . \]

\[ \therefore u_1 = \exp \left( -\frac{1}{2} r^2 \right) \quad u_2 = \theta \]

are uniform

\[ \Rightarrow g_1, g_2 \text{ Gaussian} . \]
Calculations for General Distributions

1) Example: Exponential

Suppose $x$ has distribution $e^{-x}$ for $x = 0$ to $\infty$.

$$e^{-x} \, dx = d(e^{-x})$$

$y = e^{-x}$ is uniform.

$$x = -\ln y \quad \text{where } y \text{ is a uniform random variable } 0 \leq \frac{1}{2} 1$$

2) More generally

$$p(x) \, dx = d(F(x))$$

Then $w = F(x)$ is uniformly distributed

$$w = \int_{-\infty}^{x} p(x) \, dx$$

could be different lives until

if $x$ alive $p(x) = 0 \quad x \leq 0$.

You must be able to invert integral to be able to do this

For $p(x) = e^{-x}$

$$w = 1 - e^{-x} \quad \text{is essentially as alive because } e^{-x} = 1 - w \text{ as same as } e^{-x} = y \text{ as } y \text{ and } 1 - w \text{ we both}$$
uniform between 0 and 1.

iii) Interpolation:

Suppose \( p(x) \) is given by a table

\[
\begin{array}{cc}
   x_1 & p(x_1) \\
   x_2 & p(x_2) \\
   \vdots & \vdots \\
   x_m & p(x_m) \\
\end{array}
\]

in say equally spaced intervals

\( x_k - x_{k-1} = \delta x \).

Then you form cumulative distribution

\[
\begin{align*}
   x_1 + \delta x & \quad p(x_1) & \quad F(x_1 + \delta x) \\
   x_1 + 2\delta x & \quad p(x_1) + p(x_2) \\
   \vdots & \vdots & \vdots \\
   x_{m+1} & \quad \sum_{i=1}^{m+1} p(x_i) & = 1 \text{ if normalized}
\end{align*}
\]

Then generate random number \( r \)

\( 0 \leq r \leq 1 \).

Then use inverse interpolation to find \( x \)

\[
F^{-1}(x) = r
\]

Then \( x \) is distributed according to \( p(x) \).
(iv) Accept/Reject

Then \( p_{\text{max}} = \max_x p(x) \), \( x_1 \leq x \leq x_2 \)

Let \( w_{\text{max}} \) be any number > \( p_{\text{max}} \)

Generate \( x \) uniformly in \( \mathbb{R} [x_1, x_2] \).

Generate \( r \) uniformly in \( [0, 1] \).

If \( r > \frac{p(x)}{w_{\text{max}}} \), ignore \( x \) and start again.

\[ \leq \frac{p(x)}{w_{\text{max}}} \], accept \( x \).

(v) One can combine this with Technique in (i).

Suppose \( f(x) \, dx = d\left(F(x)\right) \)

and we can invert \( F(x) \) as in (i).

Let \( w_{\text{max}} \) be any number > \( \max_x p(x) / f(x) \).

Then generate \( x \) in \( [x_1, x_2] \) distributed according to \( f(x) \) using Technique in (i).

If use (iii) with \( p(x) \Rightarrow p(x) / f(x) \)

to "correct" \( f(x) \) to \( p(x) \).

Clearly one chooses \( f(x) \) to have
a distribution "near" $p(x)$ but with a simple enough form that one can find and invert it

(2) Accept-reject is sometimes an important technique in event generation.

Suppose we generate a set $N$ of events $i$ with weight $w_i$ representing the production probability. Then if we want to calculate some simple quantity $q_i$, the rule is to take

$$<Q> = \frac{\sum_{i=1}^{N} q_i w_i}{\sum_{i=1}^{N} w_i} \quad (1)$$

An alternative is to use acceptance-rejection to find a set $M$ of events with "weight 1",

$$M < N$$

$$<Q> = \frac{1}{m} \sum_{j=1}^{m} q_j \quad (2)$$

when is (2) superior to (1).

If you are just calculating $<Q>$, then (1) usually uses all the information you have and is "best".

However if you are doing a
lot of work to calculate $q_i$, or maybe there are many different quantities $q_i$, the accept-reject technique is best. Thus it is clearly a waste to spend a lot of time calculating $q_i$ for an event whose weight $w_i$ is small e.g. $0.001 w_{max}$.
NATIONAL INSERTABLE-TAB INDEXES ENABLE YOU TO
MAKE YOUR OWN SUBJECT ARRANGEMENT, USING PLAIN
INSERTS ON WHICH TO WRITE YOUR OWN CAPTIONS.

The Beaded edge on tab makes it easy to insert captions
Made in U.S.A.
Maximum likelihood is Best! (See Brandt p138-143, EDJRS chapters 5 and 7)

Consider a Statistic \( S(x_1, \ldots, x_n) \) where \( x_i \) are independent random numbers drawn from the same distribution. Suppose we are using \( S \) to estimate some parameter \( \lambda \) characterising the distribution. We can assume that \( \langle S \rangle \sim \lambda \) (if \( \langle S \rangle \sim x^3 \) then replace \( S \) by \( \sqrt[3]{S} \) etc.)

The bias

\[
B(\lambda) = E(S) - \lambda
\]

This should be distinguished from the variance \( \sigma^2(S) \)

\[
\sigma^2(S) = \langle (S - E(S))^2 \rangle
\]

For the maximum likelihood method

\( B(\lambda) \sim \frac{1}{n} \) and \( \sigma(S) \sim \frac{1}{\sqrt{n}} \)

\[\text{Bias} \quad \text{true value of } \lambda \]

In likelihood

\[\text{Standard Deviation} \quad \text{Mean value of} \quad \text{Distribution} = E(S)\]
As the statistical error is $\sqrt{n}$ times larger than the bias, one usually ignores it. The maximum likelihood method has asymptotically small bias and the minimum possible variance. We can only bound the variance for unbiased estimators (Statistics S) as for instance, the statistic $S=0$ has zero variance but obviously a bad bias (unless $\lambda = 0$!).

Let $P(x_i, \lambda)$ be probability distribution of the $x_i$.

$$
\lambda + B(\lambda) = \int S \exp(\ln L) \, dx_1 \ldots dx_n
$$

where $L = \prod_{i=1}^{n} P(x_i, \lambda)$

Differentiating with respect to $\lambda$, we get

$$
1 + B'(\lambda) = \int S \left[ \frac{\partial}{\partial \lambda} (\ln L) \right] \exp(\ln L) \, dx_1 \ldots dx_n
$$

$$
= < S \left[ \frac{\partial}{\partial \lambda} (\ln L) \right] >
$$

as $< \frac{\partial}{\partial \lambda} (\ln L) > = \frac{\partial}{\partial \lambda} \int \exp(\ln L) \, dx_1 \ldots dx_n$

$$
= \frac{\partial}{\partial \lambda} \frac{1}{L} = 0
$$

By normalization of probabilities.
we can replace \( \bar{s} \) by \( s - \text{any constant} \) or in particular \( s - \langle s \rangle \)

\[ 1 + B'(\lambda) = \langle \left( (s-\langle s \rangle) \frac{\partial}{\partial x} (\ln x) \right) \rangle \]  
(9)

Now for any random variables \( \alpha, \beta \) we have \( \langle \alpha \beta \rangle^2 \leq \langle \alpha^2 \rangle \langle \beta^2 \rangle \)  
(\#)

This is obviously because for any constants \( a, b \)

\[ \langle (a \alpha + b \beta)^2 \rangle = a^2 \langle \alpha^2 \rangle + 2ab \langle \alpha \beta \rangle + b^2 \langle \beta^2 \rangle \geq 0 \]

This is implies (\#).

Applying this inequality to (9) we have

\[ \left( 1 + B'(\lambda) \right)^2 \leq \sigma^2(s) \left( \langle \frac{\partial^2}{\partial x^2} (\ln x) \rangle \right)^2 \]

or \( \sigma(s) \geq \frac{1 + B'(\lambda)}{\sqrt{\langle \left( \frac{\partial}{\partial x} (\ln x) \right)^2 \rangle}} \)  
(CA)

We set the denominator as \( \sqrt{I'(\lambda)} \) where the "information" \( I'(\lambda) \) is given by:
\[ I(\lambda) = \int \left[ \sum_{i=1}^{n} \frac{\partial}{\partial \lambda} (\ln P(x_i, \lambda)) \right] \left[ \sum_{j=1}^{n} \frac{\partial}{\partial \lambda} (\ln P(x_j, \lambda)) \right] L(x_1, x_2 \ldots x_n) \, dx_1 \ldots dx_n \]

The off diagonal terms are just
\[ <\frac{\partial}{\partial \lambda} (\ln P(x_i, \lambda))> <\frac{\partial}{\partial \lambda} (\ln P(x_j, \lambda))> = 0 \]
by same argument as on the bottom of ml2.

\[ \therefore I(\lambda) = n \int \left[ \frac{\partial}{\partial \lambda} (\ln P(x, \lambda))^2 \right] P(x, \lambda) \, dx \]

This can also be written as
\[ -\int \frac{d^2}{d\lambda^2} (\ln P(x, \lambda)) P(x, \lambda) \, dx \]

\[ I(\lambda) \] is information on \( \lambda \) conditioned on measurement \( x_i \). Notice that \( I \) has intuitively reasonable property (for name to be "information") of being
proportional to the number of observations. (CR) on bottom of ML3 (The Craner Rao bound) states that for an unbiased statistic

\[
\sigma(S) = \frac{1}{\sqrt{I(\lambda)}}.
\]

Relevance to Maximum Likelihood Method

The maximum likelihood statistic \( S \) is given by the equation

\[
\frac{\partial}{\partial S} \ln \xi L(S, x_1, x_2, \ldots, x_n) = 0
\]

or

\[
\sum_{L=1}^{n} \frac{\partial}{\partial S} \ln P(x_i, S) = 0. \quad (1)
\]

Expand \( \ln P(x_i, S) \) about \( S = \lambda \)

\[
\ln P(x_i, S) = \ln P(x_i, \lambda) + (S - \lambda) \frac{\partial}{\partial \lambda} \ln P(x_i, \lambda)
\]

\[
+ \frac{1}{2} (S - \lambda)^2 \frac{\partial^2}{\partial \lambda^2} \ln P(x_i, \lambda) + \ldots.
\]

we will return to this

Substitute (2) into (1), to find:
\[ S = \sum \lambda \cdot \ln P(x_i, \lambda) \]
\[ \frac{\sum \partial \ln P(x_i, \lambda)}{\partial \lambda} \]

This is of our Standard form

\[ \sum \frac{f(x_i)}{g(x_i)} \]
\[ f = \frac{\partial \ln P(x, \lambda)}{\partial \lambda} \]
\[ g = -\frac{\partial^2 \ln P(x, \lambda)}{\partial \lambda^2} \]

\[ \langle S - \lambda \rangle = \frac{\langle f \rangle}{\langle g \rangle} = 0 \quad (\langle f \rangle = 0 \text{ by } A(\star)) \]

argued on bottom of ML2).

\[ \langle (S - \lambda)^2 \rangle = \frac{1}{n} \langle (f - \frac{\langle f \rangle}{\langle g \rangle})^2 \rangle \]

\[ = \frac{\langle f^2 \rangle}{n \langle g \rangle^2} \]

but \[ \langle f^2 \rangle = \langle g \rangle = \frac{I(\lambda)}{n} \quad \text{(see middle of ML4).} \]

\[ \langle (S - \lambda)^2 \rangle = \frac{1}{I(\lambda)} \]

\[ \text{i.e. the maximum likelihood method is} \]
\[ \text{Unbiased - See } (\star) \]
\[ \text{Optimal i.e. Variance = Cramer Rao bound } (\star \star) \]

This is only true in the limit \( n \to \infty \).
Let us examine this in more detail.

**Bias in Maximum Likelihood Method for finite n**

There are two corrections to the formula \( <s-\lambda> = 0 \) derived on the previous page. The first is the truncation involved in (7).

In fact:

\[
\left[ -\sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda^2} \ln P(x_i, \lambda) \right] (s-\lambda) \\
= \sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \ln P(x_i, \lambda) + \frac{1}{2} (s-\lambda)^2 \sum_{i=1}^{n} \frac{\partial^3}{\partial \lambda^3} \ln P(x_i, \lambda)
\]

Here we can replace \((s-\lambda)^2\) by \(\frac{1}{n} I(\lambda)\)

and \(\sum_{i=1}^{n} \frac{\partial^3}{\partial \lambda^3} \ln P(x_i, \lambda)\) by \(n <\frac{\partial^3}{\partial \lambda^3} \ln P(x, \lambda)>\)

with an error of \(O(1/n)\) as the each of random variables \((s-\lambda)^2\) and \(\sum_{i=1}^{n} \frac{\partial^3}{\partial \lambda^3} \ln P(x_i, \lambda)\)

is peaked about its mean with error of \(O(1/n)\).

E.g. in our standard example

\[F \tilde{g} = (n <\tilde{g}> + \sqrt{n} \tilde{f}) (n <g> + \sqrt{n} \tilde{g})\]

with \(<\tilde{f}> = <\tilde{g}> = 0\).

\[F \tilde{g} = n^2 <f><g> + O(n \sqrt{n}) \text{ term with zer mean}
and \ O(n) \text{ term with noise mean}.\]
Consider $F/G$ with $\langle f \rangle = 0$

\[
\langle \frac{F}{G} \rangle = \left\langle \left\{ \frac{\sum_{i=1}^{n} f(x_i)}{n\langle g \rangle + \sum_{i=1}^{n} [g(x_i) - \langle g \rangle]} \right\} \right\rangle \\
\sim \frac{1}{n \langle g \rangle} \left\langle \sum_{i=1}^{n} f(x_i) \left[ 1 - \frac{1}{n \langle g \rangle} \sum_{i=1}^{n} [g(x_i) - \langle g \rangle] \right] \right\rangle \\
= -\frac{1}{n \langle g \rangle} \left\langle \sum_{i=1}^{n} f(x_i) \sum_{i=1}^{n} [g(x_i) - \langle g \rangle] \right\rangle \\
\text{irrelevant as multiplier term with zero,}
\text{only } i=j \text{ counts as } x_i \text{ independent} \\
= -\frac{1}{n^2 \langle g \rangle^2} \left\langle \sum_{i=1}^{n} f(x_i) g(x_i) \right\rangle \\
= -\frac{1}{n \langle g \rangle^2} \langle fg \rangle \iff \\
\text{i.e. when } f \text{ and } g \text{ are correlated with}
\text{we get a } O(\frac{1}{n}) \text{ correction to mean.}

Apply this to with $f = \frac{\partial}{\partial \lambda} \ln P(x, \lambda)$

$g = -\frac{\partial^2}{\partial x^2} \ln P(x, \lambda)$.
we finally get

$$B(\lambda) = \langle (S - \lambda) \rangle = \frac{n}{2\lambda^2} < \frac{\partial^3}{\partial \lambda^3} \ln P(x, \lambda) \rangle \quad (\ast)$$

$$+ \frac{n}{\lambda^2} < \frac{\partial}{\partial \lambda} \ln P(x, \lambda) \frac{\partial^2}{\partial x^2} \ln P(x, \lambda) \rangle$$

as $I(\lambda) \propto \frac{n}{\lambda}$, we see that $B(\lambda) \sim \frac{1}{\lambda}$. 

As we can be written in a different form using

$$\frac{\partial}{\partial \lambda} < \frac{\partial^2}{\partial x^2} \ln P(x, \lambda) \rangle = \frac{d}{d\lambda} \int \frac{\partial^2 \ln P(x, \lambda)}{\partial x^2} P(x, \lambda) \, dx$$

$$= < \frac{\partial^2}{\partial x^2} \ln P(x, \lambda) \rangle + < \frac{\partial}{\partial \lambda} \ln P(x, \lambda) \frac{\partial^2}{\partial x^2} \ln P(x, \lambda) \rangle$$

but I think (\ast) is more convenient.

Note that all the terms in (\ast) can be estimated from a given data sample and so bias is corrected for. As $B(\lambda) \sim \frac{1}{\lambda}$ and estimate has statistical error $\frac{1}{\sqrt{n}}$, one can in fact find $\lambda$ with a $\lambda$
with a bias \( \approx \frac{1}{\sqrt{n}} \) in a given experiment. Normally one does not do this because statistical error is of \( O(\frac{1}{\sqrt{n}}) \) and outweighs bias.

**Example:**

If \( P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \)

where \( x \) is observed and we wish to find \( \mu, \sigma \) then maximum likelihood estimates are

\[
S(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}
\]
\[
S(\sigma^2) = \frac{\sum_{i=1}^{n} (x_i - S(\mu))^2}{n}
\]

but in fact we know that \( S(\sigma^2) \) is biased and unbiased formula is \( \frac{n-1}{n} \) times \( S(\sigma^2) \). This correction is \( O(\frac{1}{n}) \) by above theorem and indeed can be calculated from it. (after we do normal extension to several a vector of theoretical parameters \( \lambda \)).
Before going on to goodness of fit tests, we will discuss non-parametric representations of data. We will restrict ourselves here to the one-dimensional case. As we will later see, the multidimensional case is often reduced to this!

A very good reference is chapter 4 of SLAC-176, 1974 notes of Tony Feldman at a CERN Computer School.

Non-parametric data representation

Given a mass spectrum,

When a parametric summary of this is to assume that the number of counts (probability) as a function of mass can be fitted
to a Gaussian plus a polynomial background.

\[ p(m) = C \exp \left[ -\frac{(m-m_0)^2}{2\sigma^2} \right] + \sum_{i=0}^{N} a_i (m-m_0)^i \]

i.e. the data is summarized by the \(4+N\) parameters \(C, m_0, \sigma, \{a_i\}\). Often this is the best approach but it is only possible if you know an appropriate form for \(p(m)\). Even in the above case, we had to guess that smooth background could be adequately parameterized by a polynomial (of degree \(N\) chosen by trial and error).

A non-parametric representation of this data is a histogram i.e.
...divide mass range into equal intervals of size $dm$ and count the number $n_i$ of counts in the $i$'th bin $m_i + (i-1)dm \Rightarrow m_i + idm$ for $i = 1 \ldots (m_2 - m_1)/dm$. As long as the bin size is small compared with the structure in the data, the numbers $n_i$ contain all the information of the data (see D4140). In a way that is clusters but free of adhoc (and possibly wrong) assumptions about the functional form of $p(m)$.

Let us now approach the problem more formally: We have measurements $\{x_i, i=1..N\}$ where the $x_i$ are independent random
variables distributed according to some unknown probability density function $p(x)$. We wish to form estimates $\hat{p}(x) = \varphi_N(x_1 \ldots x_N)$ of $p(x)$.

**G1: Histogram Method**

Define $g_i(x)$ to be 1 in $i$th bin and zero elsewhere if $g_i(x)$ is called an indicator function. i.e. if range $a \leq x \leq b$ and

$$\Delta x = (b-a)/m,$$ then

$$g_i(x) = \begin{cases} 1 & a + (i-1)\Delta x \leq x \leq a + i\Delta x \\ 0 & \text{elsewhere} \end{cases}$$

Then we define our estimate of $p(x)$ as

$$\hat{p}_{\text{Histogram}}(x) = \frac{1}{N} \sum_{j=1}^{N} g_i(x_j) g_i(x)$$
In each bin, \( \hat{P}_n \) will for large \( N \) tend to its mean = \( \int_{\text{bin}} p(x) \, dx \) with an standard deviation that this mean = \( \frac{\text{number of } x_i \in \text{bin}}{\sqrt{N}} \) \( \sqrt{\int_{\text{bin}} p(x) \, dx} \). (See argument on D33'4 for a more accurate analysis based on correct binomial distribution.)

This method has the following disadvantages:

1) As \( N \to \infty \), \( \hat{P}_n(x) \) does not \( \to P(x) \) but rather to \( \sum_{l=1}^{\infty} g_i(x) \int_{\text{bin}} p(x) \, dx \). This can only be cured by letting bin size \( \to 0 \) as \( N \to \infty \).

2) As bin size uniform some bins
We will have lots of entries (where $p(x)$ is big) and others a small number. One would like to automatically choose larger bin size where $p(x)$ is small.

$S$2 Orthogonal Functions

A generalization of $S1$ is to replace the indicator function $g_i(x)$ by general orthogonal functions $y_i(x)$

$$\int y_i(x) y_j(x) \, dx = \delta_{ij}$$

$$\hat{P}_0(x) = \sum_{i=1}^{N} \left[ \int y_i(x) p(x) \, dx \right] y_i(x)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left( \sum_j y_i(x_j) \right) y_i(x).$$

Clearly the histogram formulae are special cases of this where
$y_i(x)$ are just indicator functions. By the central limit $\hat{p}_0(x)$ has a nice limit as $N \to \infty$ 

\[
\frac{1}{N} \sum y_i(x) \to \int y_i(x) p(x) \, dx \text{ with error } \frac{1}{\sqrt{N}}
\]

although the coefficients of $y_i(x)$ are no longer statistically independent. Again $\hat{p}_0(x)$ does not $\to p(x)$ unless you take $M \to \infty$ as $N \to \infty$ so that all terms are included in $y_i(x)$ expansion of $p(x)$.

\section{Empirical Cumulative Distribution}

Define

\[
\hat{F}(x) = 0 \quad x < x_1 \\
\hat{F}(x) = i/N \quad x_i \leq x \leq x_{i+1} \\
\hat{F}(x) = 1 \quad x > x_n
\]

where $x_i$ have been ordered in
Increasing size. This is estimate from our data of the cumulative distribution $F(x) = \int_{-\infty}^{x} p(x) \, dx$

as $\hat{F}(x) = \frac{1}{N} \sum_{j=1}^{N} \Theta(x-x_j)$, we see from central limit theorem that $\hat{F}(x)$ does indeed $\rightarrow F(x)$ as $N \rightarrow \infty$.

This empirical distribution will be very useful in goodness of fit test. Now we use it to find another estimate of $p(x)$.
4. The Rosenblatt Estimator

We just define \( \hat{P}_R(y) \) as the

name derivative of \( \hat{F}(y) \).

\[
\hat{P}_R(y) = \frac{1}{2h} \left[ \hat{F}(y+h) - \hat{F}(y-h) \right] (4.1)
\]

Actually this is just number

of points in bin \( y-h \leq x \leq y+h \)

divided by bin size. So this is

really no different from histogram

method except we have a bin of

size \( 2h \) centered at each point \( x \).

Like the other estimates to be

discussed \( \hat{P}_R(y) \) \& \( \hat{P}_R(y') \) \( y \neq y' \)

are strongly correlated. (They use

many of the same \( x_j \) in their calculation whereas

neighboring bins of a histogram where
uncorrelated. One can find the mean and variance of \( \hat{P}_R(y) \) as

\[
\langle \hat{P}_R(y) \rangle = \frac{1}{2h} \left[ F(x+h) - F(x-h) \right] \quad (4.2)
\]

\[
= p(y) + \frac{h^2}{6} p''(y) + \ldots \quad (4.10)
\]

\[
\langle (\hat{P}_R(y) - p(y))^2 \rangle = \frac{p(y)}{2h^2} + \frac{h^4}{36} [p''(y)]^2 \quad (4.3)
\]

These \( \text{results} \) are obvious on writing

\[
\hat{P}_R(y) = \frac{1}{2hn} \sum_{i=1}^{2n} j_0 \left( \frac{y-x_i}{h} \right)
\]

where \( j_0(\cdot) = 1 \) for \(-1 \leq x \leq 1\)

\[= 0 \text{ otherwise.} \]

1. is just standard \( \sigma \sqrt{N} \) events we found for in histogram case.
2. comes from \( p''(y) \) term in \( \langle \hat{P}_R(y) \rangle \). Notice these \( p''(y) \) terms would be absent if we heed
calculated behaviour of $p_2(y)$ with bin integral $\frac{1}{2h} \int \frac{F(x+h) - F(x-h)}{2h} \, dx$.

we would have found similar terms for the histogram method if we had calculated it as an estimate of $p(y)$ and not the bin integral.

From (4.3) we want $h \to 0$ while $hN \to \infty$. We can minimize (4.3) for fixed $p(x)$, $p''(y)$ by choosing $h = cN^{-\alpha}$

Putting this into the right-hand side of (4.3) and requiring $\frac{\partial}{\partial c}$ and $\frac{\partial}{\partial \alpha} = 0$, we find $\alpha = -\frac{1}{5}$ (so that both terms on the left like $N^{-4/5}$ and $C = \left[ \frac{q p(y)}{2} \frac{p''(y)^2}{1} \right]^{1/5}$ (4.5)
5 Purzen Estimates

These are just like the Rosenblatt case except that $g$ defined below (4.3) is replaced by any function whose integral is 2 from $-\infty$ to $+\infty$.

Possible choices are

$$g_0(x) = \exp(-|x|)$$

or

$$g_0 = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

These generalized estimates have very similar properties to Rosenblatt's.
G6 R-th nearest neighbour

This method allows $h$ to be automatically determined and also to vary with $x$. We could apply method more generally but here we will use it on Roseblatt's method. We choose $h$ so that interval $y-h < x \leq y+h$ has exactly $k$ points (i.e. $k$ out of set of $N$ values). Here $k$ is fixed ahead of time although it will depend on $N$. This choice leads to formula

$$\hat{P}_R(y) = \frac{R(N)}{2Nh(\bar{y}, k)} \quad (6.1)$$

where $h$ is determined for each $y$ as above.
Qualitatively this method is reasonable i.e. if gets large where $p(x)$ is small and it overcomes one of the major problems with the histogram method.

One can show that

$$< \hat{p}_r(y) > = p(y) + \frac{1}{24} \left[ \frac{k(n)}{N} \right]^2 \frac{p''(y)}{p^2(y)}$$

$$< (\hat{p}_r(y) - p(y))^2 > = \frac{p^2(y)}{k(n)}$$

(6.2)

$$+ \left[ \frac{p''(y)}{24 p^2(y)} \left( \frac{k(n)}{N} \right)^2 \right]^2$$

These are just (4.2) 4.3 with $2h$ replaced by $\frac{k(n)}{N p(y)}$ as one would expect. The qualitative origin of the two terms in (6.2, 3) is just as for the Rosenblatt case.
We need \( R(N) \to \infty \) but
\[
\frac{R(N)}{N} \to 0
\]
and we can estimate best \( R(N) \) just as for \( h \) to get
\[
\frac{R(N)}{2} \bigg|_{\text{best}} = \left\{ \frac{144 p^6(x)}{|p''(x)|^2 N} \right\}^{1/5}
\]
and \( \frac{R(N)}{2} \) behaves in some way \((N^{-1/5})\) as \( h \) did in Rosenblatt case.

As discussed by Friedman, this method fails where you hit a boundary when finding \( h(y, k) \). He recommends switching to a simple Rosenblatt near the edges of a plot. Friedman also compares all the methods on a typical problem and finds the
ranking

Nearest neighbour > Rosenblatt > Parzen (gaussian \( \mathcal{N}(y_0) \)) > Histogram

The nearest neighbour method has not caught on because of:

a) complexity of calculation. (you must order \( x_i \) for quick computing)
b) correlation between estimates \( \hat{P}(y) \) at neighboring \( y \).
and data is age). Also described in Brandt (p. 160 onwards) is the Fisher-Snedecor F distribution which can be used to test if two variances are equal.

G10 Likelihood Ratio Method

We wish to test if a particular theoretical parameter \( \lambda \) has a particular value \( \lambda^{(0)} \). Now \( H^{(0)} \) is all other values of \( \lambda \). We find

\[
T = \text{Value of likelihood maximized over all values of } \lambda / \text{Value of likelihood calculated for } \lambda = \lambda^{(0)}
\]

Note that in this ratio we all the dubious undefined parts of
cancelled out.

The value of $T$ is a generalization of $\chi^2$ test ($-2 \ln T = \chi^2_{\text{minimized}}$). $\chi^2$ calculated for $\lambda = \lambda^{(0)}$ but and clearly $T \sim 1$ implies $H^{(0)}$ is acceptable. We need to be able to calculate distribution in $T$ when guess $\lambda^{(0)}$ is true value to be able to decide if a particular $T$ is probable if $H^{(0)}$ true. This calculation is not in general very easy! (except in limit of large $n$ when $\lambda$ becomes Gaussian in $\lambda$). Brandt does the calculation for a Gaussian distribution and denotes the Student's $t$ distribution.
from this more general point of view.

**Methods based on Cumulative Distribution**

These avoid "binning" necessary in $\chi^2$ method but are only possible in 1 dimension. We start with the empirical cumulative distribution

$$F(x) = \frac{1}{n} \sum_{j=1}^{n} \Theta(x_j - x)$$

we wish to test if the $x_j$ are drawn from a particular distribution $p(x)$ with

$$F(x) = \int_{-\infty}^{x} p(y) \, dy.$$  

The Smirnov-Cramer-Von Mises test forms

$$\omega^2 = \int_{-\infty}^{\infty} \left[ F(x) - F(x) \right]^2 p(x) \, dx$$
Rather remarkably, the distribution of $w^2$ is independent of $p(x)^*$ and one can show that

$$<w^2> = \frac{1}{6n}$$
$$\sigma_{w^2}^2 = \frac{4n-3}{180n^3}$$

* This is not intuitively obvious to me but can be shown because
  1) all moments of $f(x)$ only depend on $F(x)$
  2) then we can set $z = F(x)$ in integral to remove $F$ dependence (note $dz = p(x)dx$).

One can also use

$$D = \max | F(x) - \bar{F}(x)|$$
$$-\infty \leq x \leq \infty$$

For large $n$ $D$ is constant and this is discussed on p.270 of EDSRRS. Again the distribution of $D$ is independent of $F(x)$. 
G12 Rootograms

In a normal histogram, where the ith bin contains \( N_i \) events, the error is proportional to \( \sqrt{N_i} \). The vanader of the size of error with population makes it hard to judge by eye the goodness of fit of some estimate \( \hat{p}_i(x) \) to the histogram \( \hat{p}_i(x) \).

One way around this is to plot \( \bar{\alpha}_i = \sqrt{N_i} \) for

\[
\bar{\alpha}_i = \sqrt{N_i + 8N_i} - \sqrt{N_i} = \frac{8N_i}{\sqrt{N_i}} \approx 1.
\]
So a plot of the \( \alpha \) has an error independent of the number of counts and so the goodness of fit is easier to judge. Such a plot is called a "robogram". Friedman suggests plotting the residuals

\[
\hat{r}(x) = \sqrt{\hat{\beta}^2(x)} - \sqrt{\hat{\beta}_\alpha^2(x)}
\]

as being even clearer. \( \hat{r}(x) \) should have mean zero and an error independent of \( x \), i.e., if \( \hat{\beta}_\alpha(x) = c N(x) \) then standard deviation of \( \hat{r}(x) \) is \( 2 \sqrt{c} \).

We found on one experiment where we were plotting electromagnetic shower pulse heights (plotting them to predetermined forms) that we got a dynamic range \( \sqrt{P_{\text{max}}} \rightarrow \sqrt{P_{\text{min}}} \) that was visually far better than \( P_{\text{min}} \rightarrow P_{\text{max}} \) (e.g., in our case \( P_{\text{min}} \times 10 \) cts \( P_{\text{max}} \times 1000 \) cts and so a linear histogram had a dynamic range \( \sim 100 \) whereas the robogram only varied by a factor of 10).
Multivariate mapping methods

The previous techniques for non-parametric representation do not easily generalize to more than 1 dimension, i.e., if the observable $x_i$ is a vector in $d$ dimensions.

Typically one tackles vector-valued observables by forming functions $y = f(x)$ and using univariate techniques on $y$. Most histogramming packages also allow one to examine correlations (using a "scatter plot") between two such $f$'s, i.e., $y = f_1(x)$ and $z = f_2(x)$.

If $x$ corresponds to particle parameters ($f_a$, $f_b$, $f_c$ for 3 particles $a$, $b$, $c$) in a 3-particle final state.
then perhaps
\[ y = (\delta a + \delta b)^2 \]
\[ z = (\delta b + \delta c)^2 \]
and scatterplot is usually called the Dalitz plot.

One can choose the function \( f(x) \) either by intuitive insight (e.g., one uses one's physics knowledge to suggest that mass\(^2\) in the Dalitz plot are useful variables) or from statistical analysis of the actual distribution of the \( x_i \). We will only discuss the latter method here.

A) Principal Components

Here one chooses the general linear form \( y = b \cdot x \). For each \( a \), we calculate the standard deviation of the univariate distribution of \( y \).
if

\[ V_{kl} = \frac{1}{N} \sum_{i=1}^{N} x_{r}^{(i)} x_{l}^{(i)} - \left[ \frac{1}{N} \sum_{i=1}^{N} x_{r}^{(i)} \right] \left[ \frac{1}{N} \sum_{i=1}^{N} x_{l}^{(i)} \right] \]

is the usual co-variance matrix of the observations \( x^{(i)} \), then.

\[ \sigma(y) = \sqrt{a^T V a} \] (1)

The principal axis method gives the projection axis \( a \) that maximizes \( \sigma(y) \).

This clearly corresponds to the eigenvector of largest eigenvalue for \( V \). Carrying this process we deduce that the natural axes for examining our data are just the eigenvectors of the correlation matrix \( V \).

e.g. \( y \) distributions vary from

\[ y = a_{\text{max}} \cdot x \]

\[ y = a_{\text{min}} \cdot x \]

\( y \) = largest standard deviation \( \rightarrow \) Smallest.
Now the above choices are not obviously best even when one restricts oneself to the linear form. (Nonlinear choices allow even freedom of course).

B: Projection Pursuit. (Friedman and Tukey, IEEE transactions on Computers C-23, 881 (74))

The idea here is that we wish to look for separated clusters and the above principal axis method emphasized separation but not clustering. So we try to maximize

\[ I(\hat{a}) = \sigma(\hat{a}) \cdot d(\hat{a}) \]

where \( \sigma(\hat{a}) \) is given by (\#) on G37. \( d(\hat{a}) \) is to be chosen to be larger in situations like

\[ \begin{array}{c}
\mathcal{N} \\
\Rightarrow y
\end{array} \]

than for

\[ \begin{array}{c}
\Rightarrow y
\end{array} \]
Friedman and Tukey choose
\[ d(\mathbf{a}) = \sum_{i,j=1}^{N} f(r_{ij}) \]
where \( r_{ij} \) is distance from \( x_i \) to \( x_j \) projected on axis \( \mathbf{a} \)
\[ r_{ij}^2 = |\mathbf{a} \cdot (x_i - x_j)|^2 \]
and \( f \) is any smooth function of the form
\[
\begin{array}{c}
f = 0 \\
r > R
\end{array}
\]
and we search for structure of size \( R \). Then \( d \) is big if a lot of points lie within distance \( R \) of each other.

Maximizing \( I(\mathbf{a}) \) can be performed by standard non-linear maximization methods. It cannot be solved analytically as for choice of \( o(\mathbf{a}) \) as quantity to maximize.
The method can be generalized to more than one direction \( \hat{a} \). Thus one replaces \( I(\hat{a}) \) by
\[
I(\hat{a}, \hat{b}) = \sigma(\hat{a}) \sigma(\hat{b}) d(\hat{a}, \hat{b})
\]
where in \( \sigma \) one calculates
\[
\gamma_{ij}^2 = |\hat{a} \cdot (x_i - x_j)|^2 + |\hat{b} \cdot (x_i - x_j)|^2
\]
and take \( \hat{a}, \hat{b} \) to be orthogonal directions.

C) Non parametric Representation of Multivariate Data (Friedman et al. SLAC-PUB-2116, 2466).

We can generalize the above ideas to lead to a functional estimate of the multivariate probability distribution (as discussed earlier in this course for the univariate case). We don't have all the tools necessary for this yet (we need to discuss criteria for compari...
two samples from two possibly different multivariate distributions. However, we can discuss the slightly simpler problem of "smoothing" i.e. we suppose that a "response" \( y^{(i)} \) is associated with each of our random variables \( x^{(i)} \). We wish to find an estimate of the supposed functional relation:

\[
y = y(x)
\]

In the 1D case, \( y = ax + b \). There are various well-known forms of methods (least squares fit, regression analysis, splines) which we will hopefully discuss.

To generalize to the multidimensional...
case, we suppose that

\[ y = \varphi(a\cdot x) \]

For each \( a \), we apply the univariate smoothing algorithm to the points \((a\cdot x^{(i)}, y^{(i)})\). In general this will not look very good as \( y \) is multivalued \( y \) as a function of \( a\cdot x \)

so we choose \( a \) to minimize this,

i.e. to minimize scatter about the smooth i.e. \( a \) is vector that minimizes

\[ \sum_{i=1}^{n} (y^{(i)} - \varphi(a\cdot x^{(i)}))^2 \]  

(\( x \))

Again this needs a nonlinear minimizer. Calling this vector \( a^{(i)} \) we replace

\[ y^{(i)} \rightarrow y^{(i)} - \varphi_{\min}(a^{(i)}\cdot x^{(i)}) \]

and repeat procedure until the residual sum (\( x \)) is small.
§14 Minimal Spanning Trees: (MST)

See J. L. Bentley and J. H. Friedman, *IEEE Trans on Comp C-27, 97(78)
SLAC-2623 (1980).

Given a bunch of points in a multidimensional space on which a metric has been defined i.e. $d_{ij}$ is distance between $x_i$ and $x_j$. Then the MST is gotten by joining points (called nodes) by edges in such a way that:

(i) There is a connected path between any two nodes
(ii) The sum of the lengths of the edges is the smallest possible

\[ \text{e.g. in } \begin{array}{c}
\text{nodes} \\
\text{x}
\end{array} \] with distance as Euclidean metric

Then MST is:

\[ \begin{array}{c}
\text{nodes} \\
\text{x}
\end{array} \]
and is not an MST as the \(\Sigma\) lengths of connected edges is longer than optimal solution.

and is not an MST as there is no path to A

and is not an MST as there is an unnecessary edge.

It can be shown than an MST always exists and is essentially unique (i.e. only if some of the dj are equal can there be more than one MST e.g. if \(d_{ce} = d_{de}\)

then one can complete MST by joining either \(c\rightarrow e\) or \(d\rightarrow e\).
MST's have many uses that do not concern us. Obvious examples are connecting a telephone network with minimal amount of wire and of connecting electrical components in a chip. Zahn (IEEE Trans on Computers 20, 68 (1971)) discusses use of MSTs to find general patterns. Dogen (see 543) uses the MST to find jets in e+e− collisions.

\[
\begin{align*}
\text{appears as a} \\
\text{bundle of hadrons loosely clustered around original parton directions.} \\
\text{If we establish a metric} \\
\text{dy = 1/f}_{11} \text{ of particle} \\
\text{' wrt j' (where i'}
\end{align*}
\]

is that one of i} which is lower momentum and j' is higher momentum particle.
(Dorfan uses a different metric: ours is motivated by the limited use of particles until original jet exists). Then MST will naturally connect particles from the same jet and when one "jumps" from one jet to another one will use an edge which is longer than normal. Such long edges (called bridges) can be recognized and the collection of hadrons naturally decompose into jets. The same technique is used by Zahn to break a general band of points into isolated clusters (part of the "pattern recognition" problem).

We will use MSTs later in a slightly different as they provide a natural ordering of the points in a multidimensional space.

Note that Berkley and Friedman show that MSTs can be found in the typical NlogN sorting time.
105 Comparison of Multivariate Distributions

We now return to the problem discussed in 10,11 for the univariate case.

This problem has two variants.

1. Given two sets of points \( \{ x^{(1)}_{ij} \}_{i=1}^{N_1} \) \( \{ x^{(2)}_{ij} \}_{j=1}^{N_2} \), are these two sets drawn from the same probability distribution?

2. Given one set of points \( \{ x^{(1)}_{i} \}_{i=1}^{N} \), is it drawn from a given probability distribution \( p(x) \)? This second case can be reduced to the first if one uses a Monte Carlo method to generate a set of points according to \( p(x) \). In this case \( p(x) \) could be the probability distribution found from the maximum likelihood method.
$A: R^{th}$ Nearest Neighbour Comparison of two Sets

Join our two sets into a Sample of Size $N_1 + N_2$ but remember which of two sets each point comes from. Fix $R$.

Now look at each point and find the $R$ points from joint sample that are nearest to the point. Record the number $n_2(k)$ of points from set 1 that are among our $R$ nearest neighbours.

Consider the $N_1$ values $S_1: \{n_2(k)\}$ of $R$ counts among $R$th nearest neighbours of set 1 points; and $N_2$ values $S_2: \{n_2(k)\}$ of $R$ counts among $R$th nearest neighbours of set 2 points. Then if the two sets are really the same, the two integer sets $S_1$ and $S_2$ are also the same. The point is that we swapped multivariate comparison to a univariate comparison. The latter can be performed by simple modifications of the methods described in pg 11. (The latter is phrased for case of comparing sample with its supposed distribution but the modification for comparing two samples is straightforward.)

We can discuss how the integers in the sets $S_i$ should be distributed.
If the vector sets \( \{ x^{(1)}, x^{(2)} \} \) bear no resemblance to each other

\[ \text{Set 1:} \quad \text{Set 2:} \]

then nearest neighbors to a Set 1 point are all Set 1 and to a Set 2, all Set 2.

Then \( S_1 \) is \( \begin{pmatrix} R, R, \ldots, R \\ R, R, \ldots, R \end{pmatrix} \)

and in fact they agree even though the original vector sets were different and the integer sets are very dissimilar.

If the vector sets are in fact from the same distribution, then the distribution of \( Z_1(k) \) is near both
a by set 1 and set 2 point binomial corresponding to k trials with probability \( \frac{N_1}{N_1 + N_2} \) of success. This is not quite right because the points in set 1 points in different neighborhoods overlap; one would have to do a Monte Carlo calculation to find the distribution of \( n_2(k) \) for a particular multivariate distribution \( p(x) \). The binomial form is usually good and one can compare \( s_1: \{ n_2(k) \} \) for set 1 centers directly with this distribution using standard univariate methods for comparing a sample and an hypothesized distribution (91a).
B. **K-th Nearest Neighbour Comparison between a Sample Set and a Supposed Distribution** \( p(x) \) (problem 2 on p.547)

As we discussed, we can use the volume of each \( k \)-th nearest neighbour set to construct a nice univariate statistic for if we define

\[
V^{(k)} = \text{volume of } k\text{-th nearest neighbour set (taken as a sphere) added point } (i)
\]

and

\[
v(i) = \int \frac{1}{p(x')} \text{ Padeval } (x') \, dx'
\]

\[
= \frac{1}{2N} \sum_{j=1}^{N} \frac{1}{p(x(j))} \quad 0 \text{ point in volume}
\]

\[
\left( p(x) \text{ is hypothesised distribution, }\right.
\]

\[
\text{padeval is true distribution}
\]

\[
= \frac{1}{2} \sum_{k+1 \text{ points } \text{incl. center inside nn volume}} \frac{1}{p(x(0))}
\]
More precise seems an average of sphere whose volume is average of \( k \) and \( k+1 \) nearest neighbour spheres.
Now $\{V^{(n)}\}$ can be compared by our univariate methods again.

This method only tests $p$ near our sample points $x^{(n)}$; if $p$ is very large in regions where $x^{(n)}$ do not exist, then one will only indirectly see this through the normalization of $p$. This defect is not present in the alternate method of generating points according to $p(x)$ and apply the two sample comparison method.

Note that no methods are sensitive to metric used to define volume. It can be shown that it is best to transform co-ordinates so that covariance matrix is the unit matrix.
C: Methods Based on MST

Consider again two samples of size \( N_1 \) and \( N_2 \). Then the univariate Smirnov test for equality of their underlying distribution can be phrased as follows:

(i) Order the points by ascending value irrespective of the sample they came from.

(ii) For each \( i, 1 \leq i \leq N_1 + N_2 \) set

\[
    d_i = \frac{\text{Set 1 points where position in list is } i}{N_1} - \frac{\text{Set 2 points where position in list is } i}{N_2}
\]

\[
    D = \max_i |d_i|
\]

We generalize this to the multivariate case by taking our vectors \( x^n \) and constructing the Minimal Spanning Tree (MST) of the pooled data. We can use this to order the data as follows:
1) Find the "beginning (or end)" of the tree. This is found by taking each node and finding the minimum length of a path to each other node. Then the path of maximum length joins the beginning and end nodes of the tree.

2) Order the points on tree by starting at beginning and traverse tree hierarchically - choosing nearest node whenever there is a choice.

This shows a typical ranking.

3) Calculate $d_i$ and $D$ just as on page 53.
An alternate univariate method is the Wald-Wolfowitz run test. We start as in §53 (1) but simply count the number of "runs" i.e., number of consecutive sequences of the same label.

If $N_1 = 2$, $N_2 = 1$, there are three orderings of these labels: $221$, with $\text{runs} = 2$ and $\text{ordered points}$ set to which part belongs $\rightarrow \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$ $\rightarrow \frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$

values of ordered points

If the samples come from the same distribution, then the distribution of the runs can be calculated by combinatorics — independent of the distribution. The mean number of runs is

$$\begin{align*}
\text{runs} & = 1 + \sum_{k=2}^{2N} \frac{n_1}{n_2} \frac{n_1}{n_2} \frac{2^2}{2^2-1} + \frac{n_1}{n_2} \frac{n_1}{n_2} \frac{2^2}{2^2-1} \\
N & = N_1 + N_2.
\end{align*}$$

First part: $R$-th type 1; $r$-th type 2 $\Rightarrow$
by counting chance of a change in
set \( H \) at \( k \)th point of pooled data.

\[
\langle \text{runs} \rangle = 1 + \frac{2N_1N_2}{N_1+N_2} = 2 \frac{1}{3}\text{ in example on p.55.}
\]

We can clearly calculate standard
deviation const \& \( \sigma^2 = \frac{2N_1N_2(ZN_1N_2-N)}{N^2(N-1)} \)
of a distribution which becomes as
usual Gaussian for large \( N_1, N_2 \).

This method is even easier to
generalize to the multivariate case.
Take the MST as before but there is no
need to rank points. Rather remove
all edges that link points nodes,
points \( x^{(1)} \) from different samples.
Then the number of connected subtrees
that remain is exactly analogous
to the number of runs.
D: Estimate of a multivariate distribution.

Given \( \{ x^{(i)} \}_{i=1}^N \), we wish to estimate the underlying distribution \( p(x) \). Here is a possible technique:

1. Transform co-ordinates to the principal axes so that the covariance matrix of the \( x^{(i)} \) is diagonal. We invent a first guess

\[
t(x) = t_1(x_1) t_2(x_2) \ldots t_d(x_d) \quad (\star)
\]

where \( x \) is a \( d \)-dimensional vector.

The product form \((\star)\) is consistent with a diagonal covariance matrix but the latter only implies \((\star)\) if each \( t_i \) is Gaussian. We can find \( t_i(y) \) by projecting our sample onto each (principal) axis and using our (non-parametric) univariate methods.
Now (\(\tau\)) may be good enough - use the methods described in B to find \(\tau\).

If not, try

\[ P(x) = t(x) \cdot r(x) \]

and form for each point \(x^{(i)}\) in sample

\[ R = \frac{1}{k} \int_{\text{k'th nearest neighbour}} \frac{1}{t(x)} P \text{achue} (x) \, dx \]

where \(P\text{achue}\) is a trial probability distribution. If \(P\text{achue} = P(x)\)

\[ R = \frac{1}{V_k} \int_{V_k} r(x) \, dx = V_k r(x^{(i)}) \]

where we have assumed that \(r(x)\) is slowly varying over volume.

We deduce that

\[ r(x^{(i)}) = \frac{1}{V_k(x^{(i)})} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{1}{t(x^{(j)})} \]

\(k+1\) points \(x^{(i)}\) in \(k\)'th nearest neighbour volume to
Now we have associated a "response" \( r(x^{(i)}) \) to each point \( x^{(i)} \). We can use the techniques described in §13 to smooth (parameterize) \( r(x^{(i)}) \) and so we have now obtained our final estimate \( t(x) \approx r(x) \) to the underlying distribution. We did have to assume that \( r(x) \) was smooth over the nearest neighbour volume but even if it wasn't the data sample does not have the information in it to tell structure over scales small compared to this volume.

* use \( \log r \) to keep parameterisation for \( r \) positive!
5.17 Multivariate Decisions

Suppose we want to classify a vector random variable \( x \) into one of \( k \) classes (you replace two by any integer) possibilities. We discussed the univariate case in 5.8. This problem crops up often e.g. you wish to classify a particle as either \( \pi^- \) or electron (here \( x \) represents momentum measurement and pulse heights in various parts of a calorimeter). Alternatively you wish to classify a candidate track as either good or bad. Here \( x \) is a set of such things as \( x^2 \) for segments of tracks making up candidate, \# chamber hits in each segment, information on whether hits are already used in another good
track.

The general problem takes as data training vectors which are sets \( \{ x^{(1,i)} \}_{i=1}^{N_1} \), \( \{ x^{(2,i)} \}_{i=1}^{N_2} \) which are known to belong to be type I and II respectively. On the basis of these we are required to design the optimal algorithm to classify an arbitrary \( x \) as type I or II. Note that if \( \int p^{(1)}(x) p^{(2)}(x) \, dx \) is negative then there is an intrinsic ambiguity in at least some regions of \( x \) space.

**A \( k \)-th Nearest Neighbor**

Given \( x \), find the \( k \) nearest neighbors to \( x \) from the pooled data \( \{ x^{(1,i)} \}, \{ x^{(2,i)} \} \) of training vectors.
Classify \( x \) as type I/II if there are more type I/II points in this nearest neighbour set.

Comments:
(i) If \( N_1 \neq N_2 \), one should weight events, e.g., if \( N_1 > N_2 \) only count type I events as if weighted \( N_2/N_1 \).
(ii) One should (a) choose \( \mathbf{R} \) and (b) transform co-ordinates to change metric for nearest neighbour distance calculation; by requiring these choices to minimize this classification of vectors in forming samples, i.e., to minimize \( x \) of type II points within nearest neighbour volume of type I points.

I believe this method if applied carefully should work well but it is computationally quite tricky (a) in implementation alone is lengthy as nonlinear minimization (b) care needed near edges of data space in nearest neighbour
methods. (c) Well known but lengthy to find nearest neighbours.

B: Projection Method


In the univariate case, suppose types I/II have cumulative distribution $F_{i,II}(x)$. Suppose we make a simple cut $x < x^*$ as I and $x > x^*$ as type II. ($F_1(\cdot | x^*) > F_2(\cdot | x^*)$). Then given a type I event chance of misclassification is $1 - F_1(x^*)$ and for a type II chance is $F_2(x^*)$. Thus average misclassification probability is $\frac{1 - F_1(x^*) + F_2(x^*)}{2}$ and is minimized by choosing $x^*$ to maximize

$$D_2(x^*) = \left| F_1(x^*) - F_2(x^*) \right|.$$
We now describe Friedman's multivariate generalization of this.

Look at each component $x_j$ of $x$ and find $D_j = \max \left\{ \hat{F}_x(x_j), \hat{F}_x(x_j) \right\}$ where the cumulative distributions are estimated as in §3 from training samples. Choose $j$ with maximum $D_j$. From previous page, a decision based on this $j$ will have the smallest misclassification.

The cut $x \leq x_j$ will now divide sample into 2 and we can repeat procedure and so build a binary tree with each node choosing a component with the smallest misclassification. This tree terminates if either (1) a given sample only contains training events of one
type. (ii) There are less than a certain preset number k of events in Sample.

Having built tree with branching vectors one can use it to make a very fast decision on any given x as to its category. This is described for the case in a memo by J. Doftan (Mark II Internal note, 1979).