Ph129
Lecture Notes

Statistics for Physicists - Eadie, Drijard, Ross

Advanced Theory of Statistics - Kendall & Stewart ?

Handbook of the Poisson Distribution, Haight

Selected Papers on Noise and Stochastic Process - various authors - early

included famous S.O. Rice paper on noise

several RMP papers on statistical physics

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

Fuller in "Courant - Studies and Essays" [510.4 C89] p. 105

On probability problems in the theory of counters.

Stochastic Process - Parzen

Decision Theory
Students, 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
Generation of Optimal Decision Algorithms
(binary trees, Wind's magnetic field
Noise
Nonparametric representations of data.
Topics requested Sam.

Very little on how to run 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 (thru 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. (Bevington (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)

Subjects:

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.

- Binomial
- Poisson
- Gaussian

Estimation of Parameters

- Maximum Likelihood
- χ^2
- Method of Moments
- Bias / Error
- Bayes v. Non Bayes
- Robust (distribution free methods).

Minimization of Functions.

Ph129c Logistics

Teacher G. Fox
Room 207 Booth
Please call x 3765, x 6673 if want to ask questions.
Typically around 8.15am → 5.30pm

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

2) M. Buchor
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": Lebesgue measure, central limit theorem etc.

4) "Parameter Estimation": χ^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 1) to 3). Typical is

<u>Feller</u>: An Introduction to Probability Theory and its applications : vols 1, 2. This only covers 1 → 3 and is very good on 1.

<u>Hoel</u>: Introduction to Mathematical Statistics : is quite readable and covers 1) to 4).

<u>Bevington</u>: Data Reduction and Error Analysis for the Physical Sciences: This covers 2, 4 and 5. It is perhaps the best simple handbook of the

formulae. However it is naive on 4), 5) and 5) and if you start doing complicated things on χ^2/maximum liklihood/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 —

P.T. Solmitz, Ann. Rev. Nucl. Sci 14, 375 (1964)

J. Orear, UCRL-8417 (1958). "Notes on Statistics for Physicists"

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

K. McDonald CTSL Internal Report #57

Arndt and MacGregor Methods in Computational Physics - Vol 6.

Annis, Cheston and Primakoff, Rev. Mod. Phys 25, 818 (53)

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 translation of Cramer into physics applications:
W.T. Eadie, 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 enriched (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

J.W. Daniel : The approximate minimization of Functionals.

So the plan of course is:

<u>D2 : Formal Points</u> : definitions, Bayes theorem, Gaussians, Central Limit Theorem.

<u>D3 : Binomial and Poisson Distributions</u>:

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

<u>D4 Estimation of Parameters</u>

Maximum likelihood, χ^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

Goodness of Fit

Useful Books for Statistics Part of Course

G.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. (North Holland) — "Statistical Methods in Experimental Physics"
– A little Sophisticated for this course but technically excellent

W.H. Press et al. — "Numerical Recipes – The Art of Scientific Computing"
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

D2/1 Formal Definitions

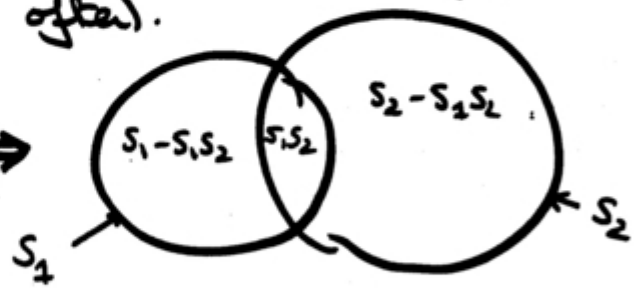
(i) A major difficulty in formulating the theory of probability is that one must relate it to the real world. Cramer (Mathematical Methods of Statistics – Princeton) amusingly notes that some of the axiomatic approaches to probability are comparable to definitions of geometry in terms of the fundamental 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 $\mathbb{R}_k$

(k dimensional real Euclidean Space). 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), $(a,b]$, $[a,b)$.. the operations of addition, subtraction or multiplication. (and continuing process denumerably often).

modern math: ⇒



$S_1 + S_2$ = 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:

i.e. $P(S_1+S_2+\ldots) = P(S_1)+P(S_2)\ldots+\ldots$

if the S_i are disjoint $S_iS_j=0$ $i\neq j$.

Note that it follows from the axioms that (a) $0\leq P(S)\leq 1$: any S. Proof: $P(S)\geq 0$ was postulated. 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 $P(S_2)\geq 0 \Rightarrow P(S_1)=P(S)\leq 1$.

(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.

Proof $S_1+S_2 = (S_1-S_1S_2)+S_2$

$P(S_1+S_2) = P(S_1-S_1S_2)+P(S_2)$ as sets disjoint. But $P(S_1-S_1S_2)\leq P(S_1)$ - 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.

(iii) 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 $\lim_{N\to\infty} N_h/N = 1/2$ where N_h is the number of times it turns up heads. We define:

$$P(\text{heads}) = \lim_{N\to\infty} N_h/N = 1/2$$

This is a random variable with only two possible values - heads or tails. Alternately we could say $x = 0$ and 1. Then in our previous language any Borel set containing 0 and not 1 has $P(S) = 1/2$, any containing 0 and 1, $P(S) = 1$ etc.

22/2 Bayes Theorem: Conditional Probabilities

Take two possible (Borel) sub-sets of $Y \in \mathbb{R}_{\mathbb{R}}$. Call them A and B.

A could be point $x=0$
B . . . $x=1$ only.

$(A+B)^*$

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

$N_1 = n(A-AB)$
$N_2 = n(B-AB)$
$N_3 = n(AB)$
$N_4 = n((A+B)^*$ - complement of $(A+B))$.

of events $n(S)$ falling in subspace S.

then $N_1+N_2+N_3+N_4 = N$. We previously defines $P(S)$ and now we expect

$$P(A) = \lim_{N\to\infty} \frac{N_1+N_3}{N}$$
$$P(B) = \lim_{N\to\infty} \frac{N_2+N_3}{N} \qquad \text{(D2.1)}$$
$$P(A+B) = \lim_{N\to\infty} \frac{N_1+N_2+N_3}{N}$$

Now we define conditional probability $p(S/A)$ {read as probability of S given A has occured} as $p(SA)/p(A)$. Note this is a probability according to axioms if we restrict Y to A.

Then in terms of frequencies.

$$p(A/B) = N_3/N_2+N_3 \quad \text{as } N\to\infty$$
$$p(B/A) = N_3/N_1+N_3 \quad \text{as } N\to\infty \qquad \text{(D2.2)}$$

whence:

$$P(B)\,P(A/B) = P(A)\,P(B/A) \qquad \text{(D2.3)}$$

which is Bayes Law. of course (D2.3) is obvious from def^n, we didn't need frequencies to prove it.

n_{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).

$p(C) = p(A)\,p(B)$ if scans A, B independent

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

$$\therefore \frac{n(C)}{n_{tot}} = \frac{n(A)}{n_{tot}} \cdot \frac{n(B)}{n_{tot}}$$

or $n_{tot} = \dfrac{n(A)\,n(B)}{n(C)}$ 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

$$p(A/B) = \frac{p(AB)}{p(B)} = p(A) \text{ independent of } B.$$

Example: page 18 of Yost

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

A = "Test Positive"

$$P(B/A) = \frac{P(B)\,P(A/B)}{P(A)}$$

Apply Seperately $B = B_1$ or B_2.

$$P(B/A) = \frac{\left[\overset{P(B_2)}{.999} \times \overset{P(A/B_2)}{.02}\,\delta(1-x) + \overset{P(B_1)}{.001} \times \overset{P(A/B_1)}{.99}\,\delta(x)\right]}{\left[P(A) = .001 \times .99 + .02 \times .999\right]}$$

$P(B_1/A)$ = "ill" $= .047$

$P(B_2/A)$ = "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 theorem is take one of them

A : to be Borel set given a set of experimental results – this is correct

B : to be a set of hypotheses e.g. a set in space of values of a theoretical parameter

e.g. we measure scattering angles θ and wish to find α when probability of observing θ is $\frac{1}{2}(1 + \alpha \cos\theta)$.

A is a set in θ space.
B is a set in α space.

Then experimentalists measure $P(A/B)$ i.e. values of θ subject to known (by the almighty and those on fourth floor) single value of α.

Bayes says remarkably

$$\text{i.e. } P(B/A) = P(A/B)\frac{P(B)}{P(A)}$$

and so this gives you what you really want – the "probability distribution of α" consequent from measurements of θ. Problem with this is that:

a) α is not a random variable – it only takes one value. answer: Our knowledge of α 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 α gained from previous (other) experiments.

c) what is $p(A)$? The probability of measurement of θ given any possible of α? Nobody knows what this is! You essentially determine it by normalizing $p(B/A)$ or equivalently note that relative probabilities

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

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

D2/3 Continuous Distributions

To be formal and reveal relics of a primaeval mathematical training: take random variables in $\mathbb{R}_1$ and consider special Borel set $\xi \leq x$: 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 to $x+dx$.

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

$$p(x) = \tfrac{1}{2}\left[\delta(x) + \delta(x-1)\right]$$

(b) $$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\tfrac{1}{2}(x-\mu)^2/\sigma^2\right] \quad (D2.4)$$

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

Note that we have ensured that

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

Also the associated distribution function is:

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

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

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-y^2)\,dy .$$

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 $q(y)$. Then probability of X lying in $[x \; x+dx]$ is $p(x)\,dx$: let $[y \; y+dy]$ be associated interval in y.

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

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

<u>So</u> $$p(x) = q(f(x))\,|f'(x)| \qquad \text{(D2.5)}$$

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

D2/5: Joint Probability Distributions

We can now consider the important extension to more than one variable. Suppose we have two random variables x and y each taking values in $\mathbb{R}_1$. Then we can conceive of a random variable z which is the ordered pair (x, y). Even if we know the probability distributions of x and y

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

The distribution of z is of course 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 priori, all we know is that r satisfies

$$\left.\begin{aligned} \int r(x,y)\,dy &= p(x) \\ \int r(x,y)\,dx &= q(y) \end{aligned}\right\}$$ These are called marginal distributions (EDJRS p.18)

(a) Example: Bayes law for densities.

Let event X be x lies in $[x, x+dx]$
" " Y " y " " $[y, y+dy]$

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

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

or $t(x|y) = r(x,y)/q(y)$ a conditional distribution (EDJRS page. 18)

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

$$s(y|x) = r(x,y)/p(x)$$

or eliminating $r(x,y)$

$$t(x/y)\,q(y) = s(y|x)\,p(x) \qquad \textbf{(D2.6)}$$

This is a density version of Bayes law.

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

when y is a known function of x, clearly $s(y|x) = \delta(y - f(x))$

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

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

(c) Independence:

If the distribution of z satisfies

$$r(x,y) = p(x)\, q(y) \qquad \text{(D2.7)}$$

then we say that x and y are statistically independent.

(d) Function of 2 variables

Suppose rather than z in $\mathbb{R}_2$, we consider $h(x,y)$ in $\mathbb{R}_1$: a ~~normal~~ conventional function of x and y. Then $w = h(x,y)$ is also a random variable with distribution $d(w)$ where

$$d(w) = \iint dx\,dy\; \delta[w - h(x,y)]\, r(x,y) \qquad \text{(D2.8)}$$

This may perhaps be proved very prettily: it is obvious if you just do one-say x-integral in (D2.8) ~~by δ~~ and use same argument in D2/4.

D2/6 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: $V_x = \langle (x - \langle x \rangle)^2 \rangle$ = 2nd central moment

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

Note
$$V_x = \langle x^2 \rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2$$
$$= \underline{\langle x^2 \rangle - \langle x \rangle^2} \qquad (D2.9)$$

a result which is used continuously.

(b) n-dimensional distribution:

The mean now becomes an n-dimensional vector

$$\langle x_k \rangle = \int \otimes dx_i \; x_k \, p(x_1 \ldots x_n)$$

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

$$M_{kl} = \int \otimes dx_i \; [x_k - \langle x_k \rangle][x_l - \langle x_l \rangle] \cdot p(x_1 \ldots x_n)$$

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

$$M_{kl} = \langle x_k x_l \rangle - \langle x_k \rangle \langle x_l \rangle \quad \text{(D2.10)}$$

Also note that M is symmetric and positive semi definite. [This follows because $y_k M_{kl} y_l$ - any y - clearly has a posi ≥ 0 integrand].

Now the diagonal terms are called variances which are ≥ 0 : their

Square roots are standard deviations again. Further

$k \neq l$: $\rho_{kl} = m_{kl} / \sqrt{m_{kk} m_{ll}}$ is ≤ 1

and is called correlation coefficient.

Note that <u>if</u> $\rho_{kl} = 0$ – 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 (D2.10) 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 m_{kl} are important asymptotically (in N).

(c) Use of moment matrix:

Often the results of an experiment on $x_1 .. x_n$ are quoted in terms of means ± 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 $\underline{y} = T\underline{x}$ (T an $n \times n$ matrix) is useful. What are errors on y?

$$\langle y_k y_l \rangle = T_{kk'} \langle x_{k'} x_{l'} \rangle T^T_{l'l}$$

$$\text{or} \quad m_y = T m_x T^T \qquad (D2.11)$$

To find diagonal terms in m_y - i.e. the errors^2 on the results 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

$x_1 = \rho_{00}$
$x_2 = \rho_{1-1}$
$x_3 = \rho_{10}$

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 latter values y were related by precisely such a linear transformation $y = Tx$ and one can no longer estimate errors.

The x_i and y_i in this example correspond to expressing angular decay

$$W(\theta,\varphi) = 1 + \sum_{i=1}^{3} x_i X_i(\theta,\varphi)$$

or $$W(\theta,\varphi) = 1 + \sum_{i=1}^{3} y_i Y_i(\theta,\varphi)$$

where y_i are linear combinations of the x_i corresponding to the

$X_i(\theta,\varphi)$ being linear combinations of the $Y_i(\theta,\varphi)$.

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

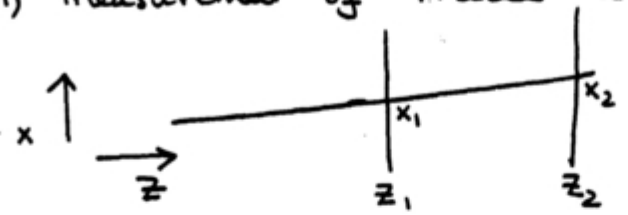
$X_1 = \cos\theta \qquad Y_1 = P_1(\cos\theta)$

$X_2 = \cos^2\theta \qquad Y_2 = P_2(\cos\theta)$

$X_3 = \cos^3\theta \qquad Y_3 = P_3(\cos\theta)$

More Examples of Correlations

1) measurement of Tracks with Chambers



Track is $x = \underset{\text{intercept}}{a} + \underset{\text{slope}}{b}\, z$

where $b = \dfrac{(x_2 - x_1)}{(z_2 - z_1)}$

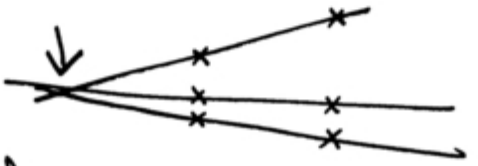
$$a = \frac{z_2 x_1 - z_1 x_2}{z_2 - z_1}$$

x_i are <u>independent</u> random variables but a and b are clearly correlated

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

where x_i has standard deviation σ_i (Note if $\sigma_1 = \sigma_2$ and $z_1 = -z_2$ i.e. origin z midpoint 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 but 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



needs it.

2) Measurement of Regge pole intercept

In E350,

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

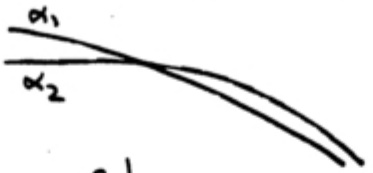
where $x \leq 1$ ($.5 \leq x \leq 1$ in experiment)

$g(t)$ 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 $\int_{.5}^{1} g(t)\,(1-x)^{1-2\alpha(t)}$ 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 terms are not given – one assumes they are zero – i.e. variables are uncorrelated. This gives:

$$\sigma_{y_i}^2 = \sum_{j=1}^{n} (T_{ij})^2 \sigma_{x_j}^2 \qquad \text{(D2.12)}$$

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

(d) Error in lack of Correlation Assumption

If $y = a_1 x_1 + a_2 x_2$

Then $\sigma_y^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + 2 a_1 a_2 \sigma_1 \sigma_2 \rho_{12}$

and all one can say in general is that:

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

The uncorrelated estimate is

$$\tilde{\sigma}_y = \sqrt{a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2}.$$

and the true σ_y satisfies

$$0 \le \sigma_y \le \sqrt{2}\,\tilde{\sigma}_y$$

where extremes are only attained when $|a_1\sigma_1| = |a_2\sigma_2|$ which is most dangerous case. Other ratios $|a_1\sigma_1|/|a_2\sigma_2| \neq 1$ are better e.g. $a_1\sigma_1 = 0 \Rightarrow \sigma_y = \tilde{\sigma}_y$!

For n variables, all one can say is

$$0 \le \sigma_y \le \sqrt{n}\,\tilde{\sigma}_y.$$

which is clearly pretty bad for large n.

(e) Parameters of a Gaussian Distribution

Suppose x is Gaussian i.e.

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

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 μ is mean.

$$\text{Variance} = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2\, p(x)\, dx$$

$y = x - \mu$ multiply and divide by σ^3.

$$= \sigma^3 / \sqrt{2\pi}\,\sigma \int_{-\infty}^{+\infty} dy\; y^2/\sigma^3 \exp\left[-y^2/2\sigma^2\right]$$

$$= \sigma^3 / \sqrt{2\pi}\,\sigma \;\; \partial/\partial\sigma \left\{ \int_{-\infty}^{+\infty} e^{-y^2/2\sigma^2}\, dy \right\}$$

$$= \sigma^2 / \sqrt{2\pi} \;\; \partial/\partial\sigma \left[\sqrt{2\pi}\,\sigma \right]$$

$$= \sigma^2 .$$

So σ is standard deviation - as its nomenclature suggested.

(f) Addition of Independent Random Variables

Suppose $y = 1/n \sum_{i=1}^{n} x_i$ where the x_i are independent and all have the same distribution with mean μ and standard deviation σ. Then clearly $\langle y \rangle$ is also μ.

$$\langle (y-p)^2 \rangle = \frac{1}{n^2} \sum_{i,j} \langle (x_i - p)(x_j - p) \rangle$$
$$= \sigma^2/n \quad \text{as all off}$$

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.

D2/7 Central Limit Theorem

(i) Moment Generating Function

This which is also called characteristic function is defined by

$$\varphi(k) = \int_{-\infty}^{+\infty} dx\, e^{ikx}\, p(x)$$

$$= 1 + ik\langle x\rangle - k^2/_{2!}\, \langle x^2\rangle + \ldots$$

(a) Note the Fourier inversion theorem gives $p(x) = {}^1/_{2\pi} \int_{-\infty}^{+\infty} dx\, e^{-ikx}\, \varphi(k)$

(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

$$\varphi_h(k) = \int_{-\infty}^{+\infty} dh\, w(h)\, e^{ikh}$$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{ik(x+y)} p(x)\, q(y)\, dx\, dy$$

$$= \varphi_x(k)\, \varphi_y(k)$$

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

(c) $\varphi(k) = \sum_n k^n \, E\left((ix)^n \right) / n!$

expectation value

$$\varphi(k) = 1 + i\langle x \rangle k - \frac{\langle x^2 \rangle k^2}{2} \quad \text{etc.}$$

For a Gaussian distribution

$$\varphi(k) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{+\infty} \exp(ixk) \exp\left[-\tfrac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right]$$

$$= \exp\left\{ i\mu k - \tfrac{1}{2} k^2 \sigma^2 \right\}$$

(d) Extended. define the Cumulant as

$$K(k) = \ln \varphi(k)$$

I have never used this!

Note that

$$K(k) = \langle x \rangle ik - \frac{\underbrace{(\langle x^2 \rangle - \langle x \rangle^2)}_{\sigma^2} k^2}{2}$$

i.e as $x \to x + p$ only shifts K by additive constant ikp, we only see "central" moments for coefficients of k^2 and above.

(ii) Central Limit Theorem

Let $y = 1/n \sum_{i=1}^{n} x_i$: x_i independent

(a) If x_i are Gaussian then so is y : in fact if y Gaussian exactly, then so must all the x_i be Gaussian.

(b) De Moivre in 1733 showed that if x_i binomially distributed, then y was always Gaussian for large n. (even though Gauss didn't exist in 1733).

(c) Laplace, in 1812, showed that de Moivre's result was generally true.

Cramer p.215. **Central Limit Theorem of Laplace**

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

Central Limit Theorem - Liapounoff

If $x_1 .. x_n$ are independent random variables with

$$\mu_i = \langle x_i \rangle$$

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

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

where ρ_i <u>must</u> exist. (not necessary in Laplace's version where all x_i had same distribution). Let

$$\rho^3 = \sum_{i=1}^{n} \rho_i^3$$

and $= \sum_{i=1}^{n} \sigma_i^2$

Then if $\lim_{n \to \infty} \rho / \Sigma \to 0$, then y is gaussianly distributed with mean $\mu = \sum_{i=1}^{n} \mu_i / n$ and S.D. Σ / n.

Note the $\lim_{n \to \infty} \rho / \Sigma$ does $\to 0$ when ρ exists in Laplace's case as $\rho \propto n^{1/3}$ and $\Sigma \propto n^{1/2}$

Last week we stated the Central Limit Theorem
Now we must prove one version of it

<u>Proof: (Laplace)</u>

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

$$\varphi_y = \left[\varphi_x(k/n)\right]^n$$

where φ_y / φ_x are c.f. of y and x respectively.

Now $\varphi_x(k) = \exp(ik\mu) \int dx\, e^{ik(x-\mu)} p(x)$

$$= \exp(ik\mu)\left[1 - \frac{k^2\sigma^2}{2} + O(k^3)\right]$$

So $\varphi_y = \exp(ik\mu)\left[1 - \frac{k^2\sigma^2}{2n^2} + O\left[(k/n)^3\right]\right]^n$

$= e^{ik\mu}\left\{\exp\left[-k^2\sigma^2/2n^2 + O\{(k/n)^3\}\right]\right\}^n$

$= \exp(ik\mu)\exp\left\{-k^2\sigma^2/2n + \underset{\substack{\downarrow \\ 0 \text{ as } n\to\infty}}{O(k^3/n^2)}\right\}$

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

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

and y is Gaussianly distributed as claimed.

(iii) Monte-Carlo Integration

Lets 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 $p(x) = 1/(b-a) \quad a \le x \le b$

$= 0 \quad$ outside $[a,b]$.

Let $\bar{I} = 1/n \sum_{i=1}^{n} f(x_i)$

where x_i are distributed according to $p(x)$. Now apply central limit theorem to $z_i = f(x_i)$.

z_i has mean $\mu = \int_a^b f(x)\,p(x)\,dx$

and $\sigma^2 = \int_a^b f^2(x)\,p(x)\,dx - \mu^2$

Then according to theorem

$$\bar{I} \to I \pm \sigma/\sqrt{n}$$

or I can be estimated as $\bar{I}$ with error $\sigma/\sqrt{n}$. Clearly a good monte carlo evaluation requires small σ. (i.e. small "variance" of f). We will return to this in numerical methods.

Note that in any Monte Carlo calculation one should calculate error in as well as value of integral.

σ involves $\int_a^b f^2(x)\, p(x)\, dx$

$$= \langle f^2(x) \rangle$$

which is estimated as

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

So one should accumulate Σf^2 as well as Σf.

(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

$$\mu_f = \langle f \rangle$$

$$\mu_g = \langle g \rangle$$

$$\sigma_f^2 = \langle (f-\mu_f)^2 \rangle$$

$$\sigma_{fg} = \langle (f-\mu_f)(g-\mu_g) \rangle$$

$$\sigma_g^2 = \langle (g-\mu_g)^2 \rangle$$

$$M = \begin{bmatrix} \sigma_f^2 & \sigma_{fg} \\ \sigma_{fg} & \sigma_g^2 \end{bmatrix} \quad \text{: moment matrix}$$

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

$$\cancel{p(f,g)}\ p(F,G) \propto \exp\left\{-\frac{n}{2}(y-\mu)^T B (y-\mu)\right\}$$

where we have lapsed into vector notation : $y = \begin{bmatrix} F \\ G \end{bmatrix}$ $\mu = \begin{bmatrix} \mu_f \\ \mu_g \end{bmatrix}$ and $B^{-1} = M$.

<u>Proof</u>: just as in ordinary central limit theorem except you consider 2nd dimensional characteristic function $\varphi(k_1, k_2)$.

<u>Corollary</u> : if F, G are defined as above, then F/G is asymptotically Gaussian with mean μ_f/μ_g and standard deviation $\langle f\rangle/\langle g\rangle \sqrt{\frac{1}{n}\left\langle \left(f/\langle f\rangle - g/\langle g\rangle\right)^2\right\rangle}$

Examples (a) $G = 1$, mean $= \mu_f$, $\sigma = \sqrt{1/n\ \sigma_f^2}$

i.e. ordinary central limit theorem

(b) $F = G$, mean $= 1$, $\underline{\sigma = 0}$.

This extension is important as it allows you to use correlations between functions calculated from the same data (observations).

<u>example</u>

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

Define random variable θ by $\theta = 1$ event passes selection criteria

$= 0$ event fails criterion.

Then $f = \theta$ $\quad \langle f \rangle = p$, say.

$g = 1.$ $\quad \langle g \rangle = 1$

The estimate of fraction is just

$$\frac{\sum_{i=1}^{N} \theta_i}{\sum_{i=1}^{N} 1}.$$

Then error in estimate is

$$\sigma = \frac{p}{\sqrt{N}} \sqrt{\langle (f/p - 1)\rangle^2}$$

$$= p/\sqrt{N} \sqrt{\langle (f^2/p^2 - 2f/p + 1)\rangle}$$

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

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

The naive estimate is

$$\langle f \rangle = 1/n \sum_{i=1}^{n} f(x_i)$$

$$\langle f^2 \rangle = 1/n \sum_{i=1}^{n} f^2(x_i) = \langle f \rangle$$

$$\sigma = \langle f^2 \rangle - \langle f \rangle^2 = p - p^2$$

error is $\sigma/\sqrt{N}$ in agreement with above.

ii) Another example:

Suppose a scattering experiment observes θ 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 $\langle P_{n_1}(\cos\theta)\rangle = \int P_{n_1}(\cos\theta)\, p(\theta)\, d\theta$

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

$\therefore$ $\langle P_{n_1}(\cos\theta)\rangle$ need to be calculated as averages we observed θ_k

$$\langle P_n(\cos\theta)\rangle = \frac{1}{N} \sum_{k=1}^{N} P_n(\cos\theta_k)$$

Even though $P_{n_1}(\cos\theta)$ and $P_{n_2}(\cos\theta)$ are orthogonal polynomials, the means $\langle P_{n_1}(\cos\theta)\rangle$, $\langle P_{n_2}(\cos\theta)\rangle$ are not uncorrelated. Both

are calculated from same set $\{\theta_k\}$ and fluctuate in a correlated fashion as observations θ_k fluctuate.

$$f = P_{n_1}(\cos\theta)$$

$$g = P_{n_2}(\cos\theta)$$

$$F = \frac{1}{N} \sum_k P_{n_1}(\cos\theta_k)$$

$$G = \frac{1}{N} \sum_k P_{n_2}(\cos\theta_k)$$

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) A third Example

Consider a histogram in which N events with weight w_i $(i=1..N)$ and x-value x_i fall.

$$\langle x \rangle = \frac{\sum_{i=1}^{N} w_i x_i = S_3}{\sum_{i=1}^{N} w_i = S_1}$$

c.f. Monte Carlo calculation of $\frac{\int x\, w(x)\, dx}{\int w(x)\, dx}$

with $f = w x \quad g = w$

the error in this calculation is:

$$\sigma = \frac{S_3}{S_1} \sqrt{\frac{1}{N} \left\langle \left(\frac{N w x}{S_3} - \frac{N w}{S_1} \right)^2 \right\rangle}$$

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

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

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

$$S_4 = \sum_{i=1}^{N} w_i^2 x_i$$

$$S_6 = \sum_{i=1}^{N} w_i^2 x_i^2 .$$

D3 Teefoefum in Las Vegas

D3/1: Binomial and Poisson Distributions

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

(i) Bernoulli Trials

(a) 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 $p+q=1$.

Afficiandos of affimative 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 out all possible outcomes as, for instance,

$$\underbrace{htht\ldots}_{n}$$

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

$$b(k,n,p) = \binom{n}{k} p^k q^{n-k} \qquad \text{(D3.1)}$$

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 (a physics graduate student in Kingsgate Hall) Smiled exactly once during whole course.

Solⁿ: 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 $p = 1/20$ $q = 19/20$

and $n = 1000$.

Thus:

Pr {Frank Nagy Smiles once and once only} =

$$\binom{1000}{1} \cdot \frac{1}{20} \cdot \left(\frac{19}{20}\right)^{999}$$

$$\approx 50\, e^{999 \log_e 19/20}$$ which isn't very big

$$\approx 50\, e^{-51}$$

$$\approx 10^{-19}$$

(ii) Poisson Limit:

(a) This is a very important special case of ~~(D3.1)~~ the binomial distribution (D3.1) which describes anything from telephone calls to events observed in particle physics. In (D3.1), take limit of $n \to \infty$ and $p \to 0$ such that $np = \lambda$ remains finite. k is also fixed

$$b(k, n, \lambda/n) = \frac{n!}{k!\,(n-k)!} \left(\lambda/n\right)^k \left(1 - \lambda/n\right)^{n-k}$$

now $\frac{n!}{(n-k)!} = n(n-1) \cdots (n-k+1) \to n^k$ as $n \to \infty$ for fixed k. (You can also get this using Stirling's formula for the

Gamma function)

Also $\left(1-\lambda/n\right)^{n-k} \approx \left(1-\lambda/n\right)^{n} \to e^{-\lambda}$

So: $b(k, n, \lambda/n) \to \frac{n^k}{k!} \lambda^k/n^k \, e^{-\lambda}$

$$= \mathbb{P}(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad \text{(D3.2)}$$

⇒ The Poisson distribution.

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

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

Solt: 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 $p = \frac{1}{20.50} = 1/1000$.

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

Let $y = \sum_{i=1}^{N} y_i$

y_i is Poisson with mean μ
N is Poisson with mean λ

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

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

as for fixed N, y is Poisson with mean $N\mu$, we find

$Pr(y = \text{integer } r)$

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

More generally we could give y_i a non Poisson distribution.

Examples are

i) Ionization - p 53 ED JRS

Particularly interesting is sensitivity to assumption that Poisson. ~~Note that~~ Sum of N Poisson ~~variables with~~ ~~me~~ each ~~with~~ Note the Poisson distribution of 4 objects with <number of objects> = λ is much broader than Ordinary Poisson of mean 4λ. 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~~ into $4 \to 6$ 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 Poission distribution x

$$\text{Mean} = \lambda \qquad \sigma^2 = \lambda$$

$$f_2(x) = \sigma_x^2 - \langle x \rangle = 0$$

Let random variable $y = 4x$

$$\langle y \rangle = 4\lambda$$

$$\langle y^2 \rangle = \langle (4x)^2 \rangle = 16 \langle x^2 \rangle$$

$$= 16 \left(\langle x^2 \rangle - \langle x \rangle^2 + \langle x \rangle \right)$$

$$= 16 \left(\sigma_x^2 + \langle x \rangle^2 \right)$$

$$= 16 \left(\lambda + \lambda^2 \right)$$

$$\therefore \quad f_2(y) = \langle y^2 \rangle - \langle y \rangle^2 - \langle y \rangle$$

$$= 16(\lambda + \lambda^2) - 16\lambda^2 - 4\lambda$$

$$= 12\lambda \quad > 0$$

i.e. y is always broader than x

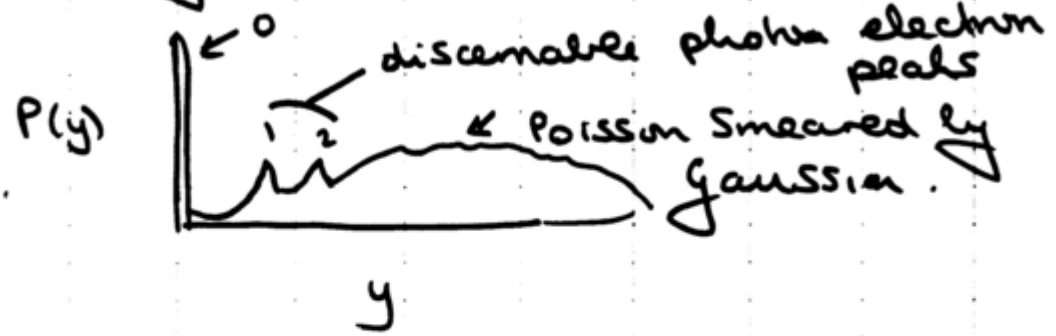
iii) Cherenkov light

Total pulse height observed

$$y = \sum_{i=1}^{N_\gamma} y_i$$

y_i = response in phototube of a single photon - probably Gaussian (i.e. limit of large Poisson)

N_γ = number of photons given off in Cherenkov medium which is certainly Poisson.



(iii) Additivity Property of Binomial or Poisson Distributions

These two distributions share an important additivity property which is α) obvious and β) leads to confusion when used in conjunction with certain

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 $X_i = k$ if $\exists$ 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 λ_1 and λ_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 λ_1, λ_2.

(iv) Gaussian limit of Binomial Distribution

(a) M. and W. 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 [redacted] 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 $y = \sum_{i=1}^{n} x_i$ is in fact, (k), the number of successes in n Bernoulli trials. Then

(α) from additivity (iii): y is Binomial with distribution $b(k,n,p)$

(β) From standard central limit theorem, y is Gaussian with mean $n\mu$ and standard deviation $\sqrt{n}\,\sigma$. Here μ and σ are mean and s.d. of $b(k,1,p)$ distribution. i.e.

$$\mu = 0.(1-p) + 1.p = p$$

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 .$$

$$= 0.(1-p) + 1.p - p^2 = p(1-p)$$

We conclude that for fixed p, 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 λ finite and let x_i be independent with each as Poisson with parameter λ. 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 $n\mu$ and standard deviation $\sqrt{n}\,\sigma$. Here again μ and σ are mean and s.d. of individual distribution.

$$\mu = \sum \exp(-\lambda)\, k\, \lambda^k / k!$$

$$= \lambda e^{-\lambda} \partial/\partial\lambda \left(\sum \lambda^k / k! \right) = \lambda e^{-\lambda} \partial/\partial\lambda (e^{\lambda})$$

$$= \underline{\underline{\lambda}}.$$

$$\sigma^2 = \langle x^2 \rangle - \lambda^2$$
$$= e^{-\lambda} \left(\lambda \partial/\partial\lambda\right)^2 (e^{\lambda}) - \lambda^2$$
$$= e^{-\lambda} \left(\lambda \partial/\partial\lambda\right) (\lambda e^{\lambda}) - \lambda^2$$
$$= \lambda + \lambda^2 - \lambda^2 = \lambda = \mu$$

So, for large Λ, any Poissonly distributed random variable y has mean Λ and standard deviation $\sqrt{\Lambda}$.

We will soon show that counts in 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 Λ = 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 t.

Consider every second: $\exists$ probability $p = N/3600$ of receiving one call: $\exists$ 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 $\lambda = N/3600$. 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

$$P_k(t) = e^{-pt} (pt)^k / k!$$ 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 ~~grows~~ continuous and not a discrete valued variable. Suppose we are studying events (e.g. telephone calls as above) that

α). Occur independently of previous history of world.

β). In time $[t, t+dt]$, the probability of an event happenning 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 to t. Then

$$P_k(t+dt) = P_k(t)\left[1-\lambda dt\right] + P_{k-1}(t)\,\lambda dt + o(dt)$$

i.e. k events at $t+dt$

or $$\frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t) \qquad : k \neq 0$$

For $k=0$, the above is modified as P_{0-1} term doesn't exist i.e.

$$dP_0(t)/dt = -\lambda P_0(t) \qquad : k=0$$

Solving the last gives

$P_0(t) = e^{-\lambda t}$ and then induction shows $P_k(t) = e^{-\lambda t}(\lambda t)^k/k!$

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.404)

Consider a system with a denumerable number of state E_n whose probability of occupation ~~population~~ varies with t. For Poisson E_n is "we have observed n events" (note only one E_n occupied at a time) and $\exists$ chance λdt of a transition $E_n \to E_{n+1}$. This is a pure birth process.

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

chance	$E_n \to E_{n+1}$	chance	$\lambda_n dt + o(dt)$
	$E_n \to E_{n-1}$	"	$\mu_n dt + o(dt)$
Any other change		"	$o(dt)$

We can at once, derive a similar

differential equation for P_n: probability of system being in state E_n.

$$dP_n(t)/dt = -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t)$$

with $\mu_0 = 0$ and $\lambda_{-1} = 0$ so that above is still correct 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 above formalism with

$\mu_n = n\mu$ μ fixed. ($1/\mu$ 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:

$$P_n(t) = e^{-\lambda/\mu} (\lambda/\mu)^n / n!$$ as

probability of n lines being busy.

It is also easy to do - the seemingly harder problem - of a finite number of telephone lines in exchange. There 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

Death = machine starts working
Birth = machine Breaks down.

and E_n = n machines broken. There is a beautiful discussion of this on pages 416 → 420 of Feller.

(f) Markov chains

One can generalize either the discrete (case (b)) or continuous ((c) to (e)) 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_{jk} or $P_{jk}\,dt$ 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 D4 refer to some additional references:

Cramer: Mathematical Methods of Statistics published by Princeton.

plus

F.T. Solmitz Ann. Rev. Nucl. Sci 14, 375 (1964)

and some old notes I have copies of

J. Orear, "Notes on Statistics for Physicists" UCRL-8417 (1958)

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

K. McDonald; CTSL Internal Report #57.

D4 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 ~~some~~ theoretical parameters describing 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 Quivering 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) darkenness) occuring during measurement, the x_i are scattered about the true value: you visit your favourite theorist (or do yourself after solving up) who believes that the probability distribution of x_i is ~~normal~~ Gaussian with mean α_0 = ~~length~~ true length and standard deviation σ_0.

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

$$p_i(x_i) = c \exp\left\{ -\tfrac{1}{2\sigma_0^2}(x_i-\alpha_0)^2 \right\}$$

The likelihood function for whole experiment of n events is defined as:

$$\mathcal{L}(x_1 \ldots x_n) = \prod_{i=1}^{n} c \exp\left\{ -1/2\sigma_0^2 (x_i - \alpha_0)^2 \right\}$$

$$= \tilde{c} \exp\left\{ -1/2\sigma_0^2 \sum_i (x_i - \alpha_0)^2 \right\}$$

This is – upto a constant – the probability of the x_i given the true length α_0 and certain theoretical assumptions we represent as T.

$$\mathcal{L}(x_1 \ldots x_n) = Pr\{ x_1 \ldots x_n \mid \alpha_0, T \}$$

in conditional probability notation. T is for instance the notion that distribution Gaussian and the value of σ_0 taken. (The latter being calculable from the strength of earthquake or amount of beer drunk).

(b) Enter Mr. Bayes

Now what we really want is the probability distribution of an estimate α of α_0. We can formally get this by use of Bayes law: first extend $p(x_i) = \exp\{-1/2\sigma_0^2 (x_i - \alpha)^2\}$ to be defined for any value of α: not necessarily α = true value α_0. Introduce:

$P(\alpha | x_i, T)$ = Probability true length is α given measurements x_i and predjudice T. Then by Bayes + or rather a slight extension of it.

$$P(\alpha | x_i, T) = P(x_i | \alpha, T) \frac{P(\alpha | T)}{P(x_i | T)} \qquad (D4.1)$$

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

This is the essence of the maximum liklihood principle. (D4.1) converts the liklihood $L(x_1 .. x_N) = P(x_i | \alpha, T)$ which is a function / Probability } of x_i for fixed

α, T into $P(\alpha | x_i, T)$ which is the desired probability distribution of α for given measurements x_i and prejudice T.

(c) Interpretation of (D4.1)

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

(β) $P(\alpha | T)$ philosophically represents a priori knowledge of α before experiment (measurement of x_i) ~~[illegible]~~ performed. This is shaky - not because we don't really know how to define what $|T$ ("given T") means - but, for instance, suppose we wish to parameterize "complete ignorance" of α before

experiment performed.

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

But suppose we put $\beta = \alpha^2$ and then β has a priori probability $g(\beta|T)$

$$g(\beta|T)\,d\beta = P(\alpha|T)\,d\alpha$$

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

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

Fortunately there is no trouble for large n because of both an obvious ~~mathematical~~ numerical reason (see next section (i)(d)) and a not so obvious mathematical reason (see section (ii)).

(d) The Liklihood Maximizes

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

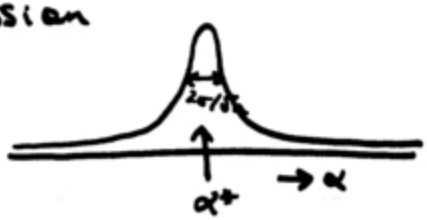
$$\mathcal{L}(x_1 \ldots x_n) = c \exp\left\{ -\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \alpha)\right\}^2$$

According to (D4.1), we now regard $\mathcal{L}$ as a function of α and so rewrite it:

$$\mathcal{L}(\alpha) = c \exp\left\{ -\frac{n}{2\sigma^2} \left(\alpha - \tfrac{1}{n} \sum_{i=1}^{n} x_i\right)^2 \right.$$

$$\left. \underbrace{- \tfrac{1}{2\sigma^2} \left[\sum_i x_i^2 - \tfrac{1}{n} \left(\sum_i x_i\right)^2\right]} \right\}$$

The last term is a constant independent of α, and so as a function of α, $\mathcal{L}$ is just a gaussian which maximizes at $\alpha = \tfrac{1}{n} \sum_{i=1}^{n} x_i$ and has width $\sigma/\sqrt{n}$.



So as $\mathcal{L} = Pr(\alpha | x_i, T)$ we say that maximum likelihood estimate of α is $\alpha^* = 1/n \sum_{i=1}^{n} x_i$ and its standard deviation is $\sigma/\sqrt{n}$.

Note that whatever $P(\alpha)$ you put in the shape of $\mathcal{L} \times P(\alpha|T)$ is essentially independent of P as $n \to \infty$ and we get a Gaussian peaked at α^* 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 $\mathcal{L}(\alpha) = \prod_{i=1}^{n} p(x_i, \alpha)$ for any

measurement x_i whose probability $p(x_i, \alpha)$ depends on value x_i and theoretical parameter(s) α.

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

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

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

For instance, this would happen if in earthquake problem you were still drunk in morning and decides to

find $\sqrt{\alpha}$. Then $\sqrt{\alpha}$ would peak at $\pm\sqrt{\alpha^*}$

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

(c) Warning Example

Consider probability distribution appropriate for golanzahen expt:

$$p(\varphi, \alpha) = (1 + \alpha\cos\varphi)/2\pi \quad : -1 \le \alpha \le 1.$$

If we had but one event

$$\mathcal{L}(\alpha) = (1 + \alpha\cos\varphi_1)/2\pi$$

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

or $\cos\varphi_1 < 0$: $\alpha = -1$ (other end of range) maximizes $\mathcal{L}$.

If you have 2 events:

$$\mathcal{L}(\alpha) = (1 + \alpha(\cos\varphi_1 + \cos\varphi_2) + \alpha^2 \cos\varphi_1 \cos\varphi_2)/(2\pi)^2$$

This has a maximum at end of range, again, unless $\cos\varphi_1 \cos\varphi_2 < 0$.

(d) So usually people warn you by saying: "If distribution of α <u>not</u> Gaussian, then do not simply quote maximum and standard deviation, rather plot the likelihood". Unfortunately this is dubious, because for <u>small</u> 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 $\mathcal{L}$ is Gaussian!

So correct warning is: "If

distribution not Gaussian – everything is OK if either:

(α) just many sharp humps as in $\sqrt{\alpha}$ example, data can be fitted with >1 value of parameter.

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

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

If (α) (β) or (γ) dont apply and α <u>isnt</u> Gaussian, then one can only be unhappy!

(iii) Lifetime Example

All books give the example of determining the lifetime τ_0 of a particle from the observation of decay times $t_1 \ldots t_n$ of n decays.

The probability distribution is Poisson:

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

Now dt is meaningless - it is independent of τ and only affects normalization of likelihood. The unknown $P(\alpha|T)$, $P(x_i|T)$ terms in (D4.1) made this arbitrary anyway!

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

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

where we lazily convert $\propto$ into an $=$ sign.

maximum is $dL/d\tau = 0$ or

$$1/\tau^{*2} \sum_{i=1}^{n} t_i - n/\tau^* = 0$$

or $\tau^* = 1/n \sum_{i=1}^{n} t_i$ 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 $y = 1/n \sum_{i=1}^{n} t_i$ is a sum over independent random variables and has mean $\langle y \rangle = \tau_0$ and standard deviation σ

$$\sigma^2 = 1/n \left\{ \langle t_i^2 \rangle - \langle t_i \rangle^2 \right\}$$

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

or $\sigma = \tau_0/\sqrt{n}$ (D4.2)

(v) Multiplication of Experiments

If we have two experiments that are independent but depend on same theoretical parameter α.

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

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

~~The~~ 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 α

24/2 χ^2 Method

This is standard and fully correct (best) method for combining results from several experiments. It is in fact used (because its easy conceptually and for computer) even when its inapplicable e.g. if there are correlations between "supposed" "independent" random variables or if the basic experiments have too few events to be Gaussian.

The problem is : given N experimental measurements x_i and errors σ_i plus estimates ξ_i of the x_i which are functions of n theoretical parameters. Then the probability of each observation is

$$p_i = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left\{ -(x_i - \xi_i)^2 / 2\sigma^2 \right\}$$

The Likelihood $L = \prod_{i=1}^{N} p_i$

$$\propto \exp\left\{ - \sum_{i=1}^{N} (x_i - \xi_i)^2 / 2\sigma_i^2 \right\}$$

So, the principle of maximizing the likelihood is equivalent to minimizing χ^2 where

$$\chi^2 = \sum_{i=1}^{N} (x_i - \xi_i)^2 / \sigma_i^2 \qquad (04/2.1)$$

wrt n theoretical parameters in ξ_i.

(i) Notice that χ^2 is the sum $\sum_{i=1}^{N} v_i^2$ where v_i are independent Gaussian random variables of unit standard deviation and zero mean. We <u>can</u> calculate χ^2 distribution exactly (M. and W. pages 393-395) but in practice this is rarely necessary. Thus we can easily show that – for fixed ξ_i – that

$$\langle \chi^2 \rangle = N$$
$$\sigma^2_{\chi^2} = 2N \qquad (D4/2.2)$$

and this is sufficient for large N as central limit theorem says χ^2 will be gaussian with above mean and standard deviation.

(ii) Now (D4/2.2) is not quite correct because if we determine the theoretical parameters to minimize χ^2 then implicitly ξ_i are also random variables that are functions of x_j. 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

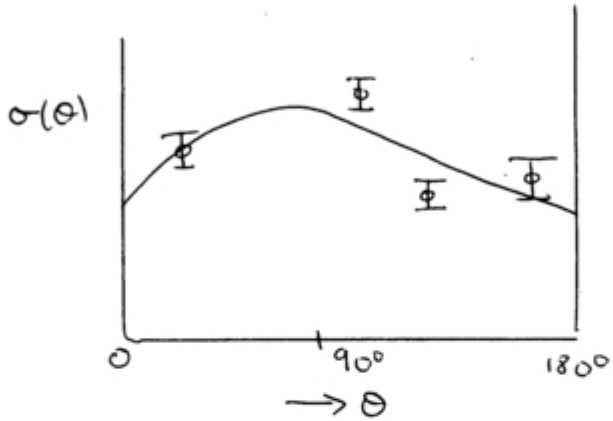
of χ^2 is many (say >2) standard deviations away from its mean, then the theoretical form is in doubt. The usual maximum likelihood method does not have this advantage: remember L was unnormalized and so its value had no significance for the goodness of fit. This is quite a difficult problem and an elementary discussion is given in Macdonald notes page 14 onwards.

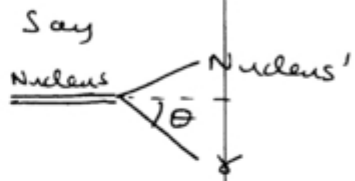
(iii) The only way to get a feeling for χ^2 is to use it: so I want dwell here on the important ~~lots simple~~ techniques used to minimize χ^2 to get good theoretical parameters and estimate errors in the same. M. and W. solve

(v) Example of use of χ^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 ξ_i are linear in unknown parameters a_j.

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

minimize : $\displaystyle\sum_{L=1}^{N} \frac{(x_i - \xi_L)^2}{\sigma_L^2}$

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

$$\sigma_i = \text{const.} \sqrt{N_i} \qquad \sigma_i = x_i/\sqrt{N_i}$$

where ith bin has N_i events.

$$\sum_{i=1}^{N} \frac{(x_i - \xi_i)}{\sigma_i^2} \cdot \frac{\partial \xi_i}{\partial a_m} = 0$$

or $\sum_{\ell} m_{m\ell}\, a_\ell^* = b_m$ * denotes χ^2 or maximum likelihood estimate

"χ^2" matrix $= \sum_i \frac{C_{im} C_{i\ell}}{\sigma_i^2}$ Positive Symmetric matrix

$$= \sum_i \frac{(\partial \xi_i / \partial a_m)(\partial \xi_i / \partial a_\ell)}{\sigma_i^2}$$

$$b_m = \sum_i \frac{x_i C_{im}}{\sigma_i^2} = \sum_i \frac{x_i}{\sigma_i^2} \frac{\partial \xi_i}{\partial a_m}.$$

$$\therefore \quad \boxed{\underline{a}^* = m^{-1} \underline{b}}$$

(vi) we can find errors in either of the two ways

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

b) Shape of likelihood.

$$\mathcal{L} = \exp\left(-\tfrac{1}{2}\chi^2\right)$$

we have – as far as shape of $\mathcal{L}$ concerned

$$\mathcal{L} = \exp\Big(\text{Independent of } a$$

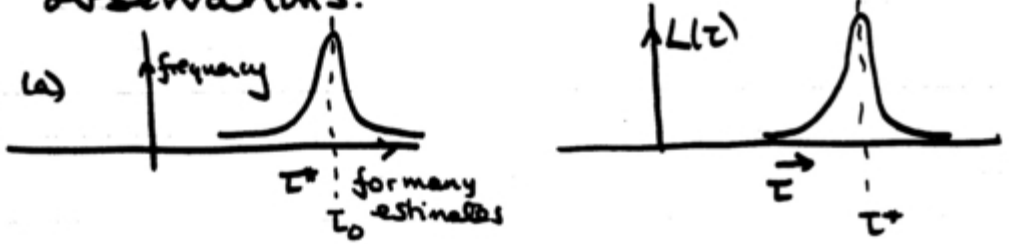
+ no terms linear in $a - a^*$ due to "maximum" condition

$$-\tfrac{1}{2}\,(\underline{a}-\underline{a}^*)^T\, M\, (\underline{a}-\underline{a}^*)\Big)$$

i.e. M is error matrix for a.

The mean and standard deviation of y refer to distribution of values of y in many experiments - each of n observations.

(b) However maximum likelihood method of finding error is, off hand, quite different. Rather than looking at distribution of y over infinitely many ~~trials~~ experiments each of n observations, it considers τ dependence of L for <u>one</u> given experiment of n observations.



Then we found it became Gaussian and calling maximum likelihood standard deviation $\tilde{\sigma}$

$$L(\tau) \propto \exp\left\{-1/2\tilde{\sigma}^2 (\tau-\tau^*)^2\right\}$$

or $$1/\tilde{\sigma}^2 = -\frac{d^2}{d\tau^2} \ln L(\tau) \Big|_{\tau=\tau^*}$$

$$= \frac{2}{\tau^3} \sum_{i=1}^{n} t_i - \frac{n}{\tau^2} \Big|_{\tau=\tau^*}$$

<u>or</u> $$\tilde{\sigma} = \tau^*/\sqrt{n} \qquad (D4.3)$$

as $\tau^* = \tau_0 + O(1/n)$ we see that (D4.2) and (D4.3) are in fact the same asymptotically (large n, that is).

Note that physics books (Solmitz is an exception) dont mention method (a) of finding error in maximum liklihood 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 <u>example</u> 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 isn't 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 calculable 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 (b) of finding error.

(iv) 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 α. We have n such observations x_i:

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

Let α^* be maximum of L wrt α.

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

$$\langle \alpha^* \rangle = \alpha_0 + O(1/n)$$

Note it is not $= \alpha_0$ exactly. (α_0 is true theoretical parameter). If for <u>any</u> estimate $\langle \alpha_{estimate} \rangle = \alpha_0$, then we say

estimate is unbiased: if $\langle \alpha_{estimate} \rangle = \alpha_0 + b(\alpha_0)$ then we have a biased estimate with bias $b(\alpha_0)$. So maximum likelihood estimate is for finite n 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 α^* as a random variable: then its standard deviation is $O(1/\sqrt{n})$ and to this order we can correctly calculate standard deviation in the "physics way" already discussed. (i.e. from α dependence of $L(\alpha)$).

(c) There is <u>no</u> other estimate of α_0 which has a smaller standard deviation to asymptotically in n i.e. maximum

likelihood method is best ~~[illegible]~~ as long as n large

<u>Proof</u>: (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 true theoretical value α_0

$$\ln L(\alpha) = L_0 + (\alpha - \alpha_0) L_1 + \frac{(\alpha-\alpha_0)^2}{2} L_2$$

$$L_0 = \sum_{i=1}^{n} \ln P(x_i, \alpha_0)$$

$$L_1 = \sum_{i=1}^{n} \partial/\partial\alpha \ \ln P(x_i, \alpha_0)$$

$$L_2 = \sum_{i=1}^{n} \partial^2/\partial\alpha \ \ln P(x_i, \alpha_0)$$

We get maximum likelihood estimate, by calculating

$$L_1 + (\alpha^* - \alpha_0) L_2 = 0$$

i.e. $\alpha^* = \alpha_0 - L_1/L_2$ (D4.4)

The error in this is $O\{(\alpha^* - \alpha_0)^2 L_3/L_2\}$

which is $O(1/n)$ as we ~~[illegible]~~ will soon show that $(\alpha^* - \alpha_0)^2$ is $O(1/n)$.

Now L_1 and L_2 are random variables and (D4.4) defines α^* as a random variable. We can calculate:

$$\langle \alpha^* \rangle = \alpha_0 - \langle L_1 / L_2 \rangle$$

$$\cancel{= \alpha_0 - \langle L_1 \rangle / \langle L_2 \rangle}$$

$$= \alpha_0 - \langle \partial/\partial\alpha \ln P(x,\alpha_0) \rangle / \langle \partial^2/\partial\alpha^2 \ln P(x,\alpha_0) \rangle$$

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

$\langle \partial/\partial\alpha \ln P(x,\alpha_0) \rangle =$ by definition

$$\int dx\, P(x,\alpha_0)\, \partial/\partial\alpha \ln P(x,\alpha_0)$$

$$= \int dx\, \partial/\partial\alpha\, P(x,\alpha_0) = \partial/\partial\alpha \left[\int dx\, P(x,\alpha_0) \right]$$

$= \partial/\partial\alpha \cdot 1$ as P is truely normalized.

$= 0$. This proves part (a) that $\langle \alpha^* \rangle = \alpha_0$ with corrections from both L_3/L_2 and $L_1/L_2 \neq$ 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 α^* as $\sqrt{-1/L_2}$ or rather this evaluated at $\alpha = \alpha^*$ not $\alpha = \alpha_0$ but these are same asymptotically.

$$\sigma^2_{physics} = -1 \Big/ \sum_{i=1}^{n} \left[\partial^2/\partial\alpha^2 \ln P(x_i, \alpha_0) \right]$$

$$\langle \sigma^2_{physics} \rangle = -1/n \langle \partial^2/\partial\alpha^2 \ln P(x_i, \alpha_0) \rangle$$

where $\langle \partial^2/\partial\alpha^2 \ln P(x, \alpha) \rangle = \int dx\, P \left\{ -\frac{(\partial P/\partial\alpha)^2}{P^2} + \frac{\partial^2 P/\partial\alpha^2}{P} \right\}$

Now second term is just $\partial^2/\partial\alpha^2 \int P$ and is zero again because of normalization condition. So

$$\langle \sigma^2_{physics} \rangle = 1 \Big/ \left\{ n \int dx\, (\partial P/\partial\alpha)^2/P \right\} \quad (D4.5)$$

we can calculate the correct <u>statistical</u> standard deviation in α^* by taking

(D4.4) and using our F/G theorem with $f_i = \partial/\partial\alpha \ln P(x_i,\alpha)$, $g_i = \partial^2/\partial\alpha^2 \ln P(x_i,\alpha)$

Generally

$$\sigma^2 = 1/n \left\langle \left(f/\langle g\rangle - g\langle f\rangle/\langle g\rangle^2 \right)^2 \right\rangle$$

but $\langle f\rangle = 0$ – proved above, so:

$\sigma^2 = 1/n \cdot \langle f^2\rangle/\langle g\rangle^2$. It is easy to see that $\langle f^2\rangle = \int (\partial P/\partial\alpha)^2/P^2 \cdot P\,dx$ and this $= \langle g\rangle$. So $\sigma^2_{\text{statistical}}$ is just $1/n\langle g\rangle$ which is exactly equal to $\langle \sigma^2_{\text{physics}}\rangle$ from (D4.5). The expectation value is no sweat: from central limit theorem $\langle\sigma^2_{\text{physics}}\rangle$ and $\sigma^2_{\text{physics}}$ differ only by relative error $O(1/\sqrt{n}) \to 0$ as $n \to \infty$.

the simplest problem i.e. ξ_i are linear in theoretical parameters. Bevington should be read before tackling a real problem.

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

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

$$x_2 = \frac{1}{n} \sum_{i=1}^{n} y_i^2 \quad \text{etc.}$$

Then they are not independent but we can use the theorem proved at the end of Section 32 to show that x_i will (for large m) be Gaussian but there is a correlation matrix

i.e. $L \propto \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\xi})^T \underline{\underline{M}} (\underline{x}-\underline{\xi})\right\}$

as discussed in Solmitz.

By diagonalizing M we can (in principle) reduce this to

the usual χ^2 method with x_i replaced by eigenvectors of $\underline{\underline{M}}$.

This idea is implicit 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 $\underline{\underline{M}}$ and forms χ^2 from non-independent random variables. This is done in high energy physics when fitting the moments $\rho_{00}, \rho_{1-1}, \rho_{10}$ 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.

vii) One should also mention that there are two versions of χ^2 depending on whether σ_i taken from "experiment or estimated theoretically. In the latter case, σ will be a function of the parameters to be varied.

For example, suppose x_1 is number of events in some counting experiment i.e. σ_i (experimental) $= \sqrt{x_1}$ but σ_1 (theory) $= \sqrt{\xi_1}$

Now x_1 is only Gaussian for large x_1 but then $\sqrt{x_1}$ and $\sqrt{\xi_1}$ are essentially same. ~~Use Cauchy-Schwarz theorem~~ If $x_1 = 0$, then theoretical σ is to be preferred! – although χ^2 method itself is in difficulties as x_1 not Gaussian.

Best to go back to maximum likelihood and put in true Poisson distribution for x_1.

D4/3 Counter Physics

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

(i) 

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 λ_i associated with them such that distribution of number n_i of events in counter is Poisson with parameter $\lambda_i t$ (t is time). This follows from our theory of Birth processes. In HEP, $\lambda_i \propto \int_{BIN} dt \, d\sigma/dt$ where BIN is t-bite covered by counter.

This is a straightforward application of likelihood:

$$p_i = e^{-\lambda_i t} (\lambda_i t)^{n_i} / n_i! \qquad (D4/3.1)$$

(n_i observed, λ_i theory : t maybe known or not).

$$\mathcal{L} = \prod_{i=1}^{m} p_i \qquad (D4/3.2)$$

$$\text{or } \ln \mathcal{L} = -\sum_{i=1}^{n} \ln n_i! - t \sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} n_i \ln(\lambda_i t)$$

(ii) Unnormalized ~~Theory~~ Experiment

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

$$\partial(\ln L)/\partial t = 0.$$

$$\text{or} \quad t = \sum_{i=1}^{m} n_i \Big/ \sum_{i=1}^{m} \lambda_i = N/\Lambda$$

where N = total number of events
Λ = Total Poisson parameter
Summed over all counters.

Now if we ~~forget~~ eliminate t, we get

$$\ln L = -\sum_{i=1}^{m} \ln n_i! - N + N \ln N + \sum_{i=1}^{m} n_i \ln(\lambda_i/\Lambda)$$

or dropping terms independent of λ_i:

$$\ln L = \sum_{i=1}^{m} n_i \ln(\lambda_i/\Lambda) \qquad (D4/3.3)$$

$$= \sum_{j=1}^{N} \ln \tilde{p}(c_j)$$

where c_j is counter where jth event added and $\tilde{p}(k) = \lambda_k/\Lambda$ 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 λ_i are functions of some theoretical parameters, we find them by maximizing (D4/3.3). In particular if all values of $p_i = \lambda_i/\Lambda$ are free parameters, we maximize (D4/3.3) subject to constraint $\sum_{i=1}^{m} p_i = 1$: this gives of course

$$p_i = n_i/N$$

(iii) Normalized Experiment

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

$$\ln L = \text{const} + \underbrace{\{N \ln(t\Lambda) - t\Lambda\}}_{(a)} + \underbrace{\sum_{i=1}^{m} n_i \ln(\lambda_i/\Lambda)}_{(b)}$$

(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 experimental observable) total counts. ($p_N = e^{-t\Lambda}(t\Lambda)^N$ which is upto to constant, usual Poisson). This requires some care to apply previous theorems as probabilities must be normalized in proofs in 84/1. However theorems do hold.

Note Orear calls this extended maximum likelihood method.

(iv) Gaussian Limit

Of course if each n_i is large

$$p_i \propto \exp\{n_i - \lambda_i t\}^2 / \lambda_i t$$

we get back χ^2 method. - note σ_i is theoretical value! (replace $\lambda_i t$ 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.

D4/4 Bubble Chamber Physics

(i) This is like D4/3 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 parameters {$(0 \le \theta \le 2\pi)$ in $(1+P\cos\theta)/2\pi$ polarization distribution} into M small bins. M should be so large that (a) probability of 2 events in a given experiment falling in same

bin is zero. (b) theoretical p_i constant over bin.

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

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

Let N be total number of events: then likelihood is

$$\mathcal{L} = \prod_{\text{successes}} p(\text{success}) \prod_{\text{failures}} p(\text{failure})$$

$$= \prod_{i=1}^{N} \exp(-p_i t\, dx_i)\ (p_i\, dx_i\, t)$$

$$\otimes \prod_{j=1}^{m-N} \exp(-p_j\, dx_j\, t)$$

working out $\mathcal{L}$ in terms of Poisson probabilities for 0 or 1 events.

Now $m \gg N$ and m bins dx_j cover space. Thus

$$\mathcal{L} = \exp\left[-\int p(x)\,dx\, t\right] \prod_{i=1}^{N} p_i\, dx_i\, t \ldots$$

Now - fortunately - arbitrary dx_i, $i = 1..N$ are independent of theoretical parameters and so can be dropped. Now (at further cost to the significance of normalization of $\mathcal{L}$).

So our fundamental formula for $\mathcal{L}$ in a continuous distribution is

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

(ii) Note - that - just as in D4/3 - if t is unknown we can find it from:

$$\partial\mathcal{L}/\partial t = 0 \quad \text{which gives}$$

$$t = N \Big/ \int p(x)\,dx .$$

when, wrt remaining parameters,

$$\mathcal{L}(\alpha) \propto \prod_{i=1}^{N} \left\{ p(x_i,\alpha) \Big/ \int p(x,\alpha)\,dx \right\}$$

- the product of normalized probability distributions.

If t 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:

$$\mathcal{L} = \exp\left\{-\int p\,dx\,t\right\} \prod_{k=1}^{K} (pt)^{n_k}$$

$$\propto \exp\left\{-\sum_{k=1}^{K} \lambda_k t\right\} \prod_{k=1}^{K} (\lambda_k t)^{n_k} \quad \text{(D4/4.1)}$$

$$\lambda_k = \int_{\text{lin for k'th group}} p\,dx$$

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 χ^2 method to analyse data.

This is the only case where χ^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 <u>throw away</u> exact values of x_i for each event - just record what bin its in. Obviously if p not constant over each bin, this throws away information. However (D4.14.1) is now exact for α and one chooses bin so that n_k large and χ^2 applicable. This is used in practice because χ^2 is so much easier and cheaper to use. Often one can bin data and minimize χ^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 piggyback further - and use a super low grade method (e.g. method of moments - see 514/5) to initialize parameters before χ^2 fit

This is discussed by Solmitz ("histogram method") and Eadie et al.

D4/5 Method of moments

(1) 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

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

where $\mu = \int p(x,\alpha)\, f(x)\, dx$.

Here μ is $\mu(\alpha)$ where α is some unknown parameter. So putting

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

one equation for α with solution $\alpha = \alpha^*$: α^* is moment estimate of α. Note :

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

(b) If we have to estimate n theoretical parameters α. Then we just get n equations by taking n moments (i.e. n different values of f).

(c) Error in μ is

theoretically $\frac{1}{\sqrt{N}}\sqrt{\int f^2\, p(x,\alpha)\,dx - \mu_0^2}$

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

using μ_0 to be true value of mean and $\mu = 1/N \sum_{i=1}^{N} f(x_i)$ 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:

$$p(x) = \tfrac{1}{2}(1+ax) \qquad -1 \le x \le 1$$

(a) ~~$\langle x^m \rangle$~~ → moment method

$$\langle x^m \rangle = \tfrac{a}{2} \int_{-1}^{+1} x^{m+1}\, dx$$

$$= \frac{a}{m+2} \qquad : m \text{ odd.}$$

So if M_m = moment method using $\langle x^m \rangle$

$$a^*_m = (m+2) \sum_{i=1}^{N} x_i^m / N$$

The error² is

$$\sigma^2\big|_m = \frac{1}{N}(m+2)^2 \left\{ \int_{-1}^{+1} x^{2m}(1+ax)\,dx - a^2/(m+2)^2 \right\}$$

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

which is a minimum for $m=1$

when $\sigma^2\big|_1 = \tfrac{1}{N}(3-a^2)$

(b) Maximum Likelihood method

Now $1/[N\sigma^2]\big|_{m.l.m.} = \int_{-1}^{+1} dx/p \, (\partial p/\partial a)^2$

– from our general theory

$$= \tfrac{1}{2} \int_{-1}^{+1} dx \; x^2/(1+ax)$$

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

$$= 1/3 + a^2/5 + a^3/7 + \cdots$$

$$= 1/3 \quad a = 0 \quad \text{which} = 1/N\sigma^2\big|_1 \text{ value}$$

$$= \infty \quad a = \pm 1.$$

So $\sigma^2\big|_{m.l.m.} = \sigma^2\big|_1 \quad a = 0$

$\qquad\qquad\qquad < \sigma^2\big|_1 \quad a \neq 0.$

(iii) 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 terms

come from $\int (f(x) - \langle f \rangle)(g(x) - \langle g \rangle) p(x)\, dx$ which is nonzero in general.

It is conventional to ignore correlations: the expansion in terms of $Y_m^l(\theta, \varphi)$ of $d\sigma/dt$ or an angular decay distribution is typical. The expansion functions Y_m^l are orthonormal, so

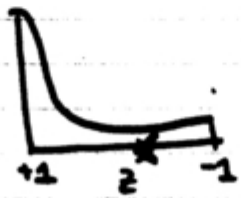
$$W(\theta, \varphi) = \sum_{lm} a_{lm} Y_{lm}$$

$$a_{lm} \propto \langle Y_{lm}^* \rangle$$

but still a_{lm} have correlated errors

as $\langle Y_{lm} Y_{l'm'} \rangle \neq \langle Y_{lm} \rangle \langle Y_{l'm'} \rangle$

These correlations in HEP example of $d\sigma/dt$



express fact that backward peak (<< forward) comes from correlated cancellations.

In formula

$$\langle f \rangle = \frac{\sum_{i=1}^{N} f(x_i)}{N} \qquad (*)$$

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 priori fixed, clearly there is no $\sqrt{N}$ error but suppose the events used in (*) are those that fall in a particular subregion (counter) of a bigger sample space (and f(x) is defined ~~over some~~ only over this subregion). Then estimate of probability of falling into counter 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 case where number of events prescribed and not subject to counting error.

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

$$P(x, y) = p(y) \; [1 + \alpha(y) x]/2$$

Take a y-bin $y_1 \le y \le y_2$

Set $S(y) = 1 \; : \; y_1 \le y \le y_2$

$= 0 \; : \;$ otherwise.

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

$$\alpha = 3 \frac{\sum_{i=1}^{N} f(x_i, y_i)}{\sum_{i=1}^{N} g(x_i, y_i)} = \frac{3F}{G}$$

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 seperately have counting errors $\sqrt{n}$ but they are correlated and correlation cancels counting errors, so that α 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 $f(x_1 \ldots x_n)$ - a function of n variables $x_1 \ldots x_n$. Equivalently we must finish the zero of an arbitary function $f'(x_1 \ldots x_n)$ of n variables. There arises in previous Section when we wish either to minimize χ^2 or maximize the likelihood $\mathcal{L}$. (equivalently to minimize $-\ln \mathcal{L}$).

The methods divide according to

(i) Can calculate f only

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

(i) 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).

D5/2 Derivative Formalism

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

$$\chi^2 = \sum_{i=1}^{N} (t_i - e_i)^2 / \sigma_i^2 \qquad (D5.1)$$

where t_i is theory - function of n parameters - we take, wolog, $n=1$ and one parameter α in the following.

$$\partial \chi^2 / \partial \alpha = \sum_{i=1}^{N} 2(t_i - e_i)\, \partial t_i / \partial \alpha \,/ \sigma_i^2$$

$$\partial^2 \chi^2 / \partial \alpha^2 = S_1 + S_2$$

$$S_1 = \sum_{i=1}^{N} 2(t_i - e_i)\, \partial^2 t_i / \partial \alpha^2 \,/ \sigma_i^2$$

$$S_2 = 2 \sum_{i=1}^{N} \left(\partial t_i / \partial \alpha \right)^2 / \sigma_i^2$$

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 $\mu = \sum_{i=1}^{N} \mu_i$ and standard deviation σ

$$\sigma^2 = \sum_{i=1}^{N} w_i^2$$

where $\frac{2(t_i - e_i)}{\sigma_i^2} \partial^2 t_i / \partial \alpha^2$ has mean μ_i and standard deviation w_i.

Now if t_i was evaluated with the theoretical parameters $\mu_i = 0$ and w_i is some number. So σ^2 is $O(N)$ and S_1 $O(\sqrt{N})$ as desired. Meanwhile S_2 is $O(N)$ as it is sum of N numbers. Whence - near solution

$$\partial^2 \chi^2 / \partial \alpha^2 = S_2 = 2 \sum_{i=1}^{N} \left(\partial t_i / \partial \alpha \right)^2 / \sigma_i^2$$

As claimed: Second derivatives of χ^2 only require evaluation of first derivatives of theory. (Only true when N large)

(ii) Lets do the same thing for maximum likelihood method.

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

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

$$\int q(x)\,dx = 1 .$$ This is crucial

but our subtle modifications in e.g. D4/3 (iii) - "generalized" maximum likelihood method can also be handled.

$$-\partial^2 \ln L / \partial\alpha^2 = S_1 + S_2$$

$$S_1 = \sum_{i=1}^{N} 1/q_i \; \partial^2 q_i / \partial\alpha^2$$

$$S_2 = \sum_{i=1}^{N} \frac{1}{q_i^2} \left(\partial q_i / \partial\alpha\right)^2$$

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.

$$\langle S_1 \rangle = N \langle 1/q \; \partial^2 q/\partial x^2 \rangle$$

$$= N \int q \cdot 1/q \; \partial^2 q/\partial x^2 \, dx$$

$$= N \; \partial^2/\partial x^2 \left\{ \int q \, dx \right\}$$

$$= N \; \partial^2/\partial x^2 \; (1) = 0 :$$

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 χ^2 and maximum likelihood fits: in the future we will only mention the former.

D5/3 The method

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

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

$$m_{jk} = \frac{1}{2} \partial^2 \chi^2 / \partial \alpha_j \partial \alpha_k = \sum_{j,k=1}^{N} \frac{1}{\sigma_i^2} \partial t_i / \partial \alpha_j \, \partial t_i / \partial \alpha_k \qquad (D5.2)$$

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

Now given any guess $\alpha = \alpha_0$

$$\chi^2(\alpha) = \chi^2(\alpha_0) + \sum_j (\alpha - \alpha_0)_j F_j + \sum_{jk} (\alpha - \alpha_0)_j (\alpha - \alpha_0)_k m_{jk}$$

Adopt a vector notation with $\Delta\alpha_j = \alpha_j - \alpha_{oj}$

$$\chi^2(\underline{\alpha}) = \chi^2(\underline{\alpha}_0) + \Delta\underline{\alpha}.\underline{F} + \Delta\underline{\alpha}^T M \Delta\underline{\alpha} \qquad (D5.3)$$

Minimum condition is

$$\partial \chi^2(\underline{\alpha}) / \partial \alpha_j = 0 \qquad \text{or}$$

$$M \Delta\underline{\alpha} = -\underline{F} \qquad (D5.4)$$

So we have new guess

$$\alpha = \Delta\alpha + \alpha_0 = \underline{\alpha}_0 - M^{-1}\underline{F}$$

for the minimum. Note that as M 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, $\underline{F} = 0$ and (D5.3) becomes

$$\chi^2(\underline{\alpha}) = \chi^2(\underline{\alpha}_0) + \Delta\underline{\alpha}^T M \Delta\underline{\alpha}$$

or remembering that the liklihood

$$\mathcal{L} = \exp(-\tfrac{1}{2}\chi^2)$$

$$\mathcal{L} = \mathcal{L}(\alpha_0) \circledast \exp\left(-\tfrac{1}{2}\Delta\underline{\alpha}^T M \Delta\underline{\alpha}\right) \quad (D5.5)$$

(D5.5) is expected Gaussian form for $\mathcal{L}$. According to D4, M is just inverse of error matrix i.e.

$$\langle(\alpha^* - \tilde{\alpha})_j (\alpha^* - \tilde{\alpha})_k\rangle = M^{-1}_{jk}$$

where $\underline{\alpha}^*$ = estimate from minimizing χ^2
$\underline{\tilde{\alpha}}$ = true parameters.

In particular "error" σ_j in α_j is

$$\sigma_j = \sqrt{(M^{-1})_{jj}} \quad (D5.6)$$

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 χ^2 change of 1.

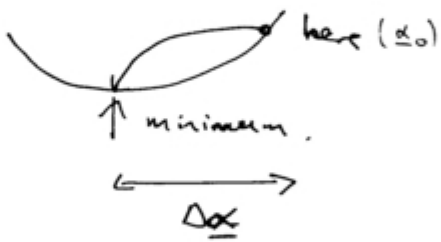
D5-4 The Real World

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

The simplest reliable algorithm is called Marquardt's method and this is intuitively easy to understand

Consider χ^2 as a function of several parameters.

χ^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 χ^2 is just a neat quadratic function and we have found a step that exactly moves to the minimum



Unfortunately, this is not usually correct as the step $\underline{\alpha}_0 \to \underline{\alpha}_0 + \Delta\underline{\alpha}$ maybe too large to support the assumption, that terms in $\dfrac{\partial^k(\chi^2)}{\partial\alpha^k}$, $k \geq 3$ are zero, to be correct

Consider the case when there are two variables.

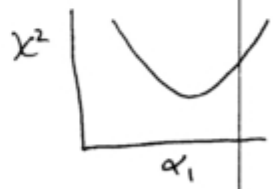
$$\chi^2 = \chi_0^2 + \text{linear terms}$$
$$+ A_1 (\alpha_1 - \alpha_1^0)^2 + A_2 (\alpha_2 - \alpha_2^0)^2$$
$$+ B_1 (\alpha_1 - \alpha_1^0)^3 + B_2 (\alpha_1 - \alpha_1^0)^2 (\alpha_2 - \alpha_2^0) + \ldots$$
$$+ \ldots$$

where we have taken the special case where there is no term proportional to $(\alpha_1 - \alpha_1^0)(\alpha_2 - \alpha_2^0)$; such a term only confuses the current discussion and we will soon see how to put it back in.

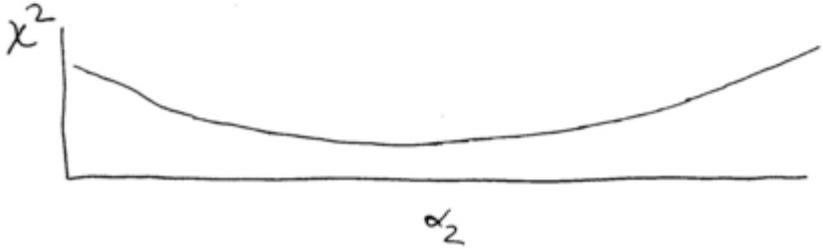
The difficulty (lack of convergence of iterative scheme) occurs when

$$A_1 >> A_2$$

So that as a function of α_1 we get



and as a function of α_2, we get



with a valley with steep walls in α_1 and gentle walls in α_2 direction.

The predicted shift from equ (D5-4), $\Delta\alpha_2 \sim 1/A_2$ and is typically large compared to $\Delta\alpha_1 \sim 1/A_1$.

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

Terms like $B_2 (\Delta\alpha_1)^2 \Delta\alpha_2$ are quite likely.

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

$$\chi^2(\underline{\alpha}_0 + \Delta\underline{\alpha}) >> \chi^2(\underline{\alpha}_0)$$

The method will diverge!

In the above example, the estimate $\Delta\alpha_1$ was valid and not spoilt by $\Delta\alpha_2$ misestimate. In a real example, there will be mixing between α_1 and α_2 due to nonzero $\frac{\partial^2 \chi}{\partial\alpha_1 \partial\alpha_2}$. Both shifts $\Delta\alpha_1, \Delta\alpha_2$ will be large and spurious.

There is an intuitively obvious solution of this. Consider the second order expansion term.

$$(\underline{\alpha} - \underline{\alpha}_0)^T M \, (\underline{\alpha} - \underline{\alpha}_0)$$

M is a symmetric positive semidefinite matrix.

M can be diagonalized

$$M \rightarrow P^T D P \quad \text{— } n \times n \text{ matrix.}$$

$$D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}$$

λ_j are eigenvalues
$\underline{P}$ are eigenfunctions of m

Put $\underline{\beta} = \underline{P}\,\underline{\alpha}$

Then $\chi^2 = \chi_0^2$ + linear terms

$$+ \lambda_1 (\beta_1 - \beta_{01})^2 + \lambda_2 (\beta_2 - \beta_{02})^2 + \cdots$$

Now the shifts $\Delta\beta_j$ can be calculated independently and do not affect each other.

Assume that we order the eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

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

$$\Delta\beta_j = \text{Predicted value } j \leq L$$
$$= 0 \quad j > L.$$

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

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

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

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

$$M = P^T \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_L \\ 0 & & & & 0 \end{bmatrix} P$$

I note that in using this method, I got to look at values of the λ_j. It was not uncommon for

$$\lambda_{max}/\lambda_{min} = \lambda_1/\lambda_n \sim 10^7 \quad \text{for } n \sim 30$$

You can ask, why one gets such large ratios. Intuitively the eigenvectors of M with small eigenvalue correspond to linear

combinations of α_j that are badly determined so very small λ'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 $\sum_{j=1}^{n} \alpha_j P_{j-1}(\cos\theta)$ 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 top index n but in general no such simple solution exists.

Marquardt's method is a simpler technique but with similar goals. One replaces

$$M \rightarrow M + \varphi I$$

where I is the identity matrix.

So means we replace $\mathcal{M}$ by

$$M = P^T \begin{bmatrix} \lambda_1 + Q & & & \\ & \lambda_2 + Q & & \\ & & \ddots & \\ & & & \lambda_n + Q \end{bmatrix} P$$

Clearly if $\lambda_L \sim Q$, then the two formalism's are highly equivalent

Eigenvectors with eigenvalues $> Q$, are shifted by naive prediction while those with eigenvalue $\ll Q$, have their shift drastically reduced.

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

. The natural "scale" for Q is given by $\frac{1}{n} \mathrm{Tr}\, M = \frac{1}{n} \sum_{j=1}^{n} \lambda_j$, the average eigenvalue.

Marquardt's method is less precise 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 fermilab events where in each event I fitted known shower shapes to measured photon energy measurements in a segmented lead-scintillator sandwich. Here $n \sim 10$ and the large number of fits (10^6) required a fast method.

§7 χ^2, Bartlett's S function

When the maximum liklihood method reduces to χ^2, it is relatively easy to judge goodness of fit.

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

where t_i depend on m theoretical parameters. Then the result of minimizing (*) wrt m theoretical parameters should be distributed according to χ^2 distribution with n−m degrees of freedom. This function is plotted on p. 64 of EDJRS. If k is number of degrees of freedom.

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

$$\langle \chi^2 \rangle = k$$

and standard deviation is $\sqrt{2k}$.

Consider ~~to~~ χ^2 with one degree of freedom

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

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

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

as $\pm(x-t)$ give same value of χ^2

$$= \frac{\exp(-\chi^2/2)}{\sqrt{2\pi}\; 2\sqrt{\chi^2}} \cdot 2$$

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

One can derive χ^2 distribution for k degrees of freedom in two ways

i) Mathews and Walker: direct calculation as above noting that $\sum_{i=1}^{k} x_i^2$ x_i mean 0, std. dev. 1 is r^2 where r is radial distance in k dimensional space. Then factors multiplying $\exp(-\chi^2/2)$ in (*) are just volume element in k dimensions

ii) Eadie et al. Show characteristic function of χ^2 with one degree of freedom is

$$\int_0^{+\infty} \exp(it\chi^2)\, p(\chi^2)\, d(\chi^2) = 1/\sqrt{1-2it}$$

∴ Characteristic function for χ^2 with

k degrees of freedom is $(1-2it)^{-k/2}$ and this can be inverse fourier transformed to give (*).

Introduce the characteristic function $F(\chi^2)$ by

$$F(\chi^2) = \int_0^{\chi^2} p(\chi^2)\, d(\chi^2) = \frac{1}{2\Gamma(k/2)} \int_0^{\chi^2} (u/2)^{(k/2)-1} e^{-u/2}\, du$$

F has probability distribution $w(F)$

$$w(F) = p(\chi^2)\, d(\chi^2)/dF = 1$$

So F is uniformly distributed between 0 and 1.

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

$C(\chi^2) = 1$ is $\chi^2 = 0$ "good"

$= 0$ is $\chi^2 = \infty$ "bad" or "low confidence"

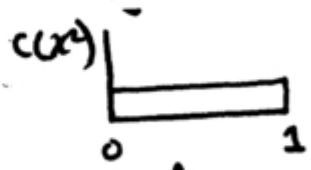
Suppose you conduct a series of experiments which gives you a sequence of χ^2 values

$\chi^2|_{1,2} \ldots$

An example in high energy physics would come from analysing a set of events each containing particle tracks

- each track would have an individual χ^2. A given experiment would generate many million such χ^2 values!

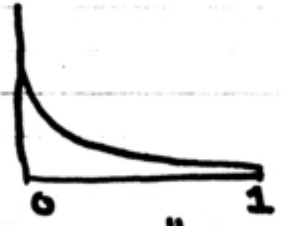
If one can do the "errors" right, the resultant distribution in $C(\chi^2)$ should be flat



but if you have "background" i.e. events that do not satisfy claimed hypothesis, then one will have an excess of events at large $\chi^2 \gg k$ or small $C(\chi^2)$

$C(\chi^2)$ will ~~be~~ look like

The plot of $C(\chi^2)$ is a good way of estimating background. Unfortunately one often gets $C(\chi^2)$ plots that look like



with no clear "flat" piece. This maybe because your "Signal" is very small. It could be because you got your errors wrong (!) e.g. maybe the true σ_i is are

all larger than ones used in calculating χ^2. In this case it makes sense to scale χ^2 ($\chi^2_{new} = c\,\chi^2_{old}$) until we get a nice flat distribution for $C \sim 1$ and analyse as on top of G19.

If we expand χ^2 near its minimum - with for convenience - just one theoretical parameter.

$$\chi^2 = \chi_0^2 + \alpha(\lambda - \lambda_0)^2 \qquad (**)$$

where λ_0 is our estimate of λ. Then probability distribution of λ is just

$$\propto \exp(-\chi^2/2)$$

$$\propto \exp\left(-\frac{\alpha(\lambda-\lambda_0)^2}{2}\right)$$

i.e. $1/\sqrt{\alpha}$ is standard deviation of estimates. Returning to (**), we find that if we ~~[illegible]~~ put $\lambda = \lambda_0 \pm 1/\sqrt{\alpha}$ (i.e. at its one standard deviation value), χ^2 is increased by 1. This χ^2 increase "rule" is common method of finding "error" in estimate λ even when the Taylor expansion (**) is not a good approximation.

This can be generalized to Bartlett's S function which is defined as

$$S(\lambda) = \frac{1}{\sqrt{I(\lambda)}} \frac{\partial \ln L}{\partial \lambda}$$

$I(\lambda)$ is information $\left\langle \left(\partial/\partial\lambda \ln L \right)^2 \right\rangle$

$$I(\lambda) = N \underbrace{\int \left[\partial/\partial\lambda \ln P(x,\lambda) \right]^2 P(x,\lambda)\, dx}_{\text{Can be calculated from theoretical forms.}}$$

N — Number of events.

From (**) ~~what that is~~ we have

$$\partial/\partial\lambda \ln L = -\frac{2\alpha}{2}(\lambda-\lambda_0) \qquad \left(L = \exp(-\chi^2/2)\right)$$

and using our earlier asymptotic result $\alpha = I(\lambda)$, we find

$$S(\lambda) = \sqrt{I(\lambda)}\,(\lambda_0-\lambda)$$
$$= \sqrt{\alpha}\,(\lambda_0-\lambda)$$

$\therefore$ $\lambda = \lambda_0 \pm 1$ standard deviation is just $S(\lambda) = \mp 1$. In fact for any n, $S(\lambda)$ has mean 0 and variance 1 and so the rule $S(\lambda) = \pm 1$ ($\pm l$ for l standard deviations) is an appropriate generalization of that to increase χ^2 by 1 (l).

§8 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 (= hypothesis that data described by claimed theory with parameters found from maximizing $\mathcal{L}$) 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 $(x_1 \ldots x_n)$ on which to make our decision. We do this by constructing yet another statistic Y which is some function of $x_1 \ldots x_n$. We assume that Y has distribution $p_Y^{(i)}$ 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 best of all possible worlds, one can find a Y such that the values taken by it under two hypotheses are disjoint i.e.

$$\int p_Y^{(0)}(y)\, p_Y^{(1)}(y)\, dy = 0 \qquad (*)$$

Unfortunately (*) is not always attaineable but one should always try to minimize the probability overlap integral. We define r as region ~~of~~ in Y Space where we will deem ~~Y to be~~ hypothesis (1) to be true and (0) to be false. ~~Rest~~ ~~Th~~ In the remaining region of Y Space, R-r, (R total Y range) we will take (0) to true, (1) to be false. This Scheme is characterized by 2 constants:

α = <u>level</u> of <u>Significance</u> or <u>Size</u> of test $= \int_r p_Y^{(0)}(y)\,dy$

α is chance that (0) will be

declared false when it is in fact true.

$\beta = \int_{R-r} p_y^{(1)}(y)\,dy$ is the contamination

i.e. chance that (0) will be declared true when it is in fact false. Clearly it depends on the application ~~wh~~ for the balance between efficiency $1-\alpha$ and contamination β.

§9 Student's t distribution

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

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

$H^{(1)}$ is mean $= \mu^{(1)}$

Define $t = \dfrac{(\bar{x} - \mu^{(i)})\sqrt{n}}{\bar{\sigma}}$

where $\bar{x} = 1/n \sum_{i=1}^{n} x_i$

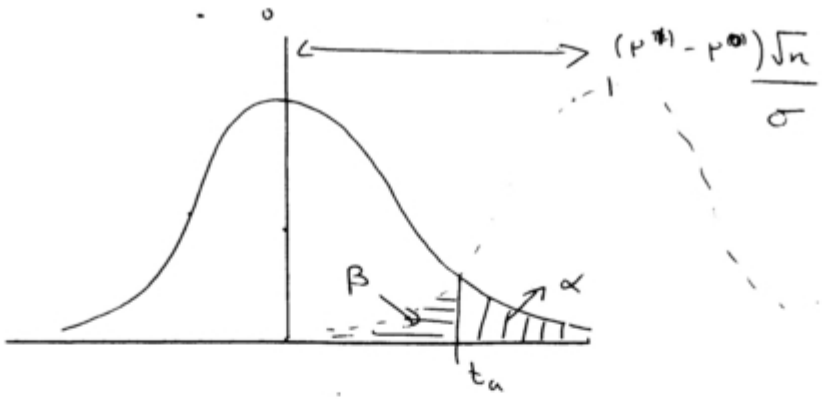
$$\bar{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$p(t) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\sqrt{n\pi}\,\Gamma(f/2)} \frac{1}{\left(1 + t^2/f\right)^{(f+1)/2}}$$

$f = 1$ 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 trivially calculable from cumulative probability functions of the Students t distribution. As discussed on page 168 of Brandt, this method can be adapted (use average $\bar{\sigma}$). to test if 2 sets of measurements have the same mean (e.g. Set 1: Smokers Set 2: non Smokers)

§16 Robust Estimation

See Andrews et al. "Robust Estimates of Location", Princeton University Press
"Robustness in Statistics, edited by R.L. Launer, G.N. Wilkinson, Academic Press (1979).

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 <u>incorrect</u> theoretical form of probability is large. Even when we know main functional dependence (Breit Wigner for a resonance, say) there are often "small" contributions to the

probability distribution (e.g. background

which must be estimated as polynomial/phase space)

which must be parameterized in an ad hoc fashion. It is hard to evaluate distortion due to ~~effect of~~ an incorrect parameterization of these unknown contributions. There is thus a great deal of interest in the study of so called <u>robust</u> 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
unhard problems that physicists are want 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 instructive study of a very simple problem; give a univariate sample $\{x^{(i)}\ i=1,N\}$ from some distribution, find an estimate of the mean μ of the underlying probability distribution. The ~~simple~~ natural method is to estimate μ from the arithmetic mean

$$\mu^*_{Am} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

As long as the underlying distribution has a finite standard deviation this will have an error $\propto 1/\sqrt{N}$ which

$\to 0$ as $N \to \infty$. Further the method is unbiased for any N

i.e. $\langle \mu^*_{AM} \rangle = \mu$.

If the $x^{(i)}$ are distributed according to either the gaussian

$$p(x) \propto \exp\left(-(x-\mu)^2/2\sigma^2\right)$$

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 $y = x^2$ where x had an exponential distribution. Then the best estimate of the mean of

y is obviously

$$y^*_{clear} = \left(\frac{1}{N} \sum_{i=1}^{N} (y^{(i)})^{1/2} \right)^2$$

and <u>not</u> $y^*_{Am} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)}$

The problem of estimating μ is hardest for distributions $p(x)$ that have ~~the~~ 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

$$\mu^*_{trim} = \sum_{i=pN+1}^{N(1-p)} x^{(i)} / N(1-2p)$$

where pN events are left off each end. ~~As long as p is fixed < 0.5,~~ For the special case of distributions ~~μ^*_{trim}~~ symmetric about their mean, the trimmed μ^* will also have mean μ and error $\propto 1/\sqrt{N}$. ~~For fixed~~ The variance of μ^*_{trim} will be larger than μ^*_{Am} for Gaussianly distributed $x^{(i)}$ but as you investigate distributions broader than Gaussians you will find the variance of μ^*_{trim} becoming smaller than that of μ^*_{Am}. A special case of μ^*_{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 $z^{(i)} = \dfrac{x^{(i)} - \mu}{\sigma}$

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

$$\partial/\partial_\mu \ln \mathcal{L} = 0$$

$$\text{or} \quad \sum_{i=1}^{N} \partial/\partial_\mu \ln p\left(\frac{x^{(i)} - \mu}{\sigma}\right) = 0$$

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

$$\sum_{i=1}^{N} \psi\left(\frac{x^{(i)} - \mu}{\sigma}\right) = 0 \qquad (*)$$

where ψ is an <u>odd</u> function of its argument.

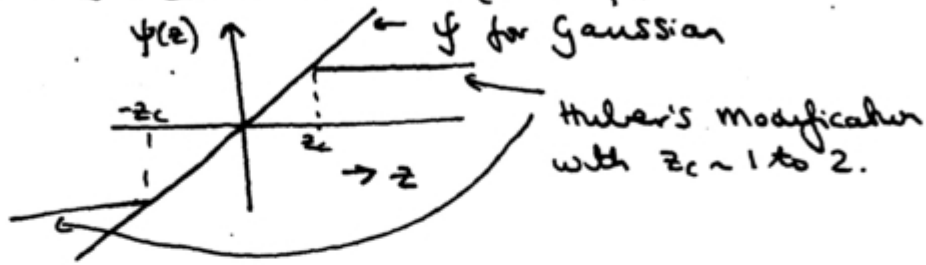
For the case of Gaussian p, ψ is just $1/\sigma^2 \left(x^{(i)} - \mu\right)$ and $(*)$ is

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} \left(x^{(i)} - \mu_\ell^*\right) = 0$$

which gives the arithmetic mean as the maximum likelihood estimate

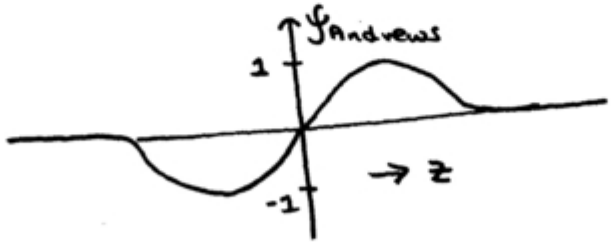
of the mean. Now we invent a function ψ which is less sensitive to $x^{(i)}$'s far from the mean than $(x^{(i)} - \mu)$.

e.g.



Many forms for ψ have been proposed. The most successful was that due to Andrews with

$$\psi(z) = \sin(z/2.1) \qquad |z| < 2.1\pi$$
$$= 0 \qquad |z| \geq 2.1\pi$$



Note z is $\left(\frac{x-\mu}{\sigma}\right)$ where σ supplies natural scale for problem. We do not need a very good estimate for σ but

it should be robust! A simple choice is σ = median $|x^{(i)} - \mu|$ which is applied iteratively as value of μ is refined.

A large number of tests are ~~defi~~ described in first reference to this section.

Let N = Gaussian Distribution - $\sigma = 1$
C = Cauchy
M = Gaussian + uniform random variable between 0 and 1.

Here is a sample of their results for 10 observations

Method	N	C	M ← Distribution
Mean	1	27314	5255
Median	1.37	3.66	7.33
Andrews	1.08	4.66	9.29

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

broad distributions. The median/Andrews methods are very good in these cases ~~but~~ with the Andrews being better in Gaussian case (σ only 4% more than best estimate) but less robust in other examples.

Note that all the methods <u>except</u> 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.

Ph 129 — 1990

G. Fox

MONTE CARLO INTEGRATION

This expands discussion on page 23 (Section (iii) of Section 1.7) of type written notes

E5/4 Monte Carlo Integration

(i) As we will find out late 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

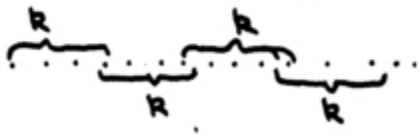
E5/4 Monte Carlo Integration

(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

Error $\propto$	$1/n^2$	iterated Trapezoidal
	$1/n^4$	iterated Simpson
	$1/n^{2m}$	iterated m-order Gauss



k = 2 Trapezoidal.
k = 3 Simpson.
k = m Gauss.

with iterated groups of k points

Taking $g(x) = 1$ for convenience, we write

$$I = (b-a)\int_{-\infty}^{+\infty} f(x)\, p(x)\, dx$$

$$p(x) = 1/(b-a) \qquad a \le x \le b$$

$$= 0 \qquad \text{outside } [a, b]$$

Now consider the Monte Carlo which

writes $I = \langle \tilde{f} \rangle$

$$\langle \tilde{f} \rangle = \int_a^b \tilde{f}(x)\, dx / b-a \qquad \tilde{f} = (b-a) f$$

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

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{f}(x_i) \to I \qquad (E5.11)'$$

with error $\sigma/\sqrt{n}$.

(E5.11)' 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 $\to 0$ 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 $\propto 1/n^{2/k}$ which has a slower n dependence as k increases. However Monte-Carlo method still keeps its $1/\sqrt{n}$ behaviour and for $k \geq 5$, 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 formulae 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 dimension. You can pose 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 feeble 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

$$I = \int_{x^2+y^2 \le 1} g(x,y)\,dx\,dy.$$

One-way of doing this is to put

$$g'(x,y) = g(x,y) \qquad x^2+y^2 \le 1$$
$$= 0 \qquad x^2+y^2 > 1$$

and calculate

$$I = \int_{-1}^{1} dx \int_{-1}^{1} dy \; g'(x,y)$$

which is an elementary Monte-Carlo integral with x and y generated independently between -1 and 1.

Actually the brute force method (generate x uniformly between -1 and 1, y uniformly between $\pm\sqrt{1-x^2}$) is more efficient - in the language of (5.11) 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 σ. Using definition of σ^2, (or remembering its name "variance") it follows that method is best when $f(x)$ is slowly varying over domain. [for instance if $f(x) \equiv$ constant, $\sigma = 0$ and method is exact). Importance Sampling ("sampling where integrand important") is a technique for reducing variation of $f(x)$.

Thus if $\quad I = \int f(x)\,dx$

Change variables $\quad y = q(x)$

$$dy = q'(x)\,dx$$

put $\quad f(x)/q'(x) = F(y)$

Then $\quad I = \int F(y)\,dy$.

We now apply Monte Carlo method generating y uniformly.

We try to choose $q'(x)$ so $f(x)/q'(x)$ 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 $\propto q'(x) = f(x)$

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 perfomed.

Example

$$f(x) = e^{-x} \otimes \text{a mess}$$

where mess slowly varying

choose $q'(x) = e^{-x}$

$$q(x) = 1 - e^{-x}$$

Then $F(y)$ = "a mess"

One can apply Monte Carlo methods to sums as well as continuous integrals.

e.g. Suppose we have

$$\int dx_a\, dx_b\, dx_c \cdots \sum_{l=1}^{2} f_l(x_a, x_b, x_c \ldots)$$

Then generate

x_a with random number

x_b

x_c

Generate another random number

r $\quad 0 \le r \le 1$ uniformly

if $(r < 0.5)$ $l = 1$

else $l = 2$

and obvious modifications for other discrete summations

Accept/Reject method

We wish to generate x according to $p(x)$ $\quad x_1 \le x \le x_2$

$$p_{max} = \max_x p(x)$$

Let w_{max} be $\geq p_{max}$ (any number)

Generate x uniformly in $[x_1\ x_2]$

" r " in $[0\ 1]$

if $r > \dfrac{p(x)}{w_{max}}$ ignore x and start again

$\le \dfrac{p(x)}{w_{max}}$ 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_t in the neighbourhood of x_i uniformly.

In the $\int p(\underline{x})\, d\underline{x}$

if $p(x_t) > p(x_i)$, accept it, and set $x_{i+1} = x_t$

if $p(x_t) < p(x_i)$, generate a random number r uniformly in $[0,1]$,

if $r < p(x_t)/p(x_i)$ set $x_{i+1} = x_t$
$r \geq$ " " $x_{i+1} = x_i$

Eventually this produces a set of points distributed "according to $p(x)$".

reasons:

1) The neighbourhoods $\Delta x - x_i \le x_t \le x_i + \Delta x$ eventually cover the whole space

2) Take any two points x_i, x_j

Suppose $p(x_i) > p(x_j)$

$$Pr(i \to j) \text{ is } \frac{p(x_j)}{p(x_i)}$$

$$Pr(j \to i) \text{ is } 1$$

i.e $$\frac{Pr(i \to j)}{Pr(j \to i)} = \frac{p(x_j)}{p(x_i)}$$

In "equilibrium"

$$N_i \, Pr(i \to j) = N_j \, Pr(j \to i)$$

$$\therefore \frac{N_i}{N_j} \propto p(x_i)/p(x_j)$$

So we have generated x_i according to $p(x_i)$

i.e. exact importance sampling according

Philosophy of Standard Experimental Monte Carlo

One is evaluating integrals of the form

$$I(\Theta) = \int \Theta(\underline{x})\, p(\underline{x})\, d\underline{x}$$

Here vector $\underline{x}$ is in parameter space of experimental events e.g. in high energy physics $\underline{x}$ labels momenta of final state particles; $p(\underline{x})$ is probability that a given $\underline{x}$ is formed e.g. in high energy physics, $p(x)$ is square of modulus of production amplitude. $\Theta(\underline{x})$ is a particular experimental observable e.g. energy into a calorimeter of a particular geometry.

Typically $p(\underline{x})$ is fixed by the theory but one wishes to calculate I for several choices of Θ as one explores different facets of the

experiment. Thus although the result of any given I depends on σ of product Θp, to get good results for many different Θ's ~~requires~~ suggests that we minimize the (effect of) variance of $p(\underline{x})$. There are three methods

1) use "importance" sampling: i.e. change variables to $\underline{y}$ where the Jacobian $\frac{\partial(\underline{x})}{\partial(\underline{y})} = J(x)$

$d\underline{x} = J(x)\, d\underline{y}$ 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 $W = \max_{\underline{x}} [p(\underline{x})]$

Then having generated $\underline{x}$ uniformly choose a random number r between 0 and 1

if $r < \frac{p(x)}{W}$ keep $\underline{x}$

$r > \frac{p(x)}{W}$ reject it.

~~Be~~ As events are accepted with probability $\propto p(\underline{x})$, they are "distributed according to $p(\underline{x})$" ie

i.e. $I(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \Theta(\underline{x}_i)$: $\underline{x}_i$ chosen as alive.

This method is particularly good if it is very hard to calculate $\Theta(\underline{x})$. This happens say if response of particular part of apparatus is hard to calculate. (It might involve tracking particles 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 $p(\underline{x})$ 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 $\langle p(x) \rangle / W$ is kept and this can be small (.01 is quite typical in problems I have worked on).

Lattice Monte Carlo

A typical "theoretical" calculation in condensed matter or high energy physics involves integrals of form

$$I = \int \exp[-H/kT] \; \Theta \; ds_1 \ldots ds_M$$

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 $10^4 \to 10^6$ dimensions are so must be evaluated by Monte Carlo

The Hamiltonian H is proportional to size of system and so the weighting function $\exp[-H/kT]$ is very small and one must be very careful to chose one integration points appropriately.

Assume you have a particular choice $s_1^{(0)} \ldots s_M^{(0)}$ with

$$H^{(0)} = H(S_1^{(0)} \ldots S_m^{(0)})$$

Then we try to move a small way from $S_i^{(0)}$ by just changing one of 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 $\Delta H \sim kT$ 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 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 $r = \exp[-[H(\ldots S_k^{(1)} \ldots)/kT - H(\ldots S_k^{(0)} \ldots)/kT]]$

if $r > 1$ (energy decreased), accept $S_k^{(1)}$

if $r < 1$ (energy increased), accept $S_k^{(1)}$ with probability r.

(b) Heath Bath – Sometimes one can use "exact" method to generate a variable according to a given probability distribution i.e. if distribution function

$$F_k(x) = \int ds_k \exp(-H(\ldots s_k \ldots)/kT)$$

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

$$\langle \Theta \rangle = \frac{1}{\substack{\#\text{ Spin} \\ \text{Configurations} \\ = N}} \sum_{l=1}^{N} \Theta(s_1^{(l)} \ldots s_m^{(l)})$$

If we had not been careful about generating the spin values

but say. chosen them uniformly, then

$$\langle \Theta \rangle = \frac{\sum_{\ell=1}^{N} \exp[-H[S^{(\ell)}]/kT] \; \Theta(S^{(\ell)})}{\sum_{\ell=1}^{N} \exp[-H[S^{(\ell)}]/kT]}$$

This is correct but numerically hopeless. The wide variation in exponential (remember H proportional to size) would give a huge standard deviation in $\langle \Theta \rangle$.

Generation of Random Numbers

See

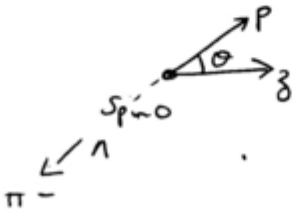
Knuth "The Art of Computer Programming" Volume 2. "Semi Numerical Algorithms"

Press "Numerical Recipes : The Art of Scientific Computing" – chapter 7

T.W. Chiu Computer Physics Communications Vol 47, 129 (1987).

Knuth is a reasonable pedagogical description although it is a little out of date – Chiu describes newer algorithms which are important in some cases. Press implements algorithms at about the level of Knuth.

One needs specialized hardware to generate true random numbers e.g.



the angle θ is random for successive decays of a bunch of π's! There are a lot of difficulties here; not

only must you measure θ in an unbiased fashion (impossible), to find a random number with 32 bits of precision would require several seperate θ values - maybe each gives 4 bits of information and so 8 values are needed!

Rather one usually a deterministic method which gives "essentially" random numbers. These 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

$$I_{j+1} = [a I_j + c] \bmod (m)$$

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 less. One could generate a cycle if

$$I_{J_1} = I_{J_2} \qquad J_1 > J_2$$

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 π

$$\pi = \int_{-1}^{+1}\int_{-1}^{+1} dx\,dy\;\theta[1-x^2-y^2]$$

then random numbers after the m'th are no longer independent and so ~~our~~ our use of central limit theorem breaks down.

Our error is $\propto \max\left[\underbrace{1/\sqrt{N}}_{\text{number of events}},\ \underbrace{1/\sqrt{m}}_{\text{cycle length}}\right]$

The references give various formal criteria to guide the choice of a c and m. Unfortunately I dont see any of these references as being very good as they dont test and compare some key choices.

In my group we have used

$m = 2^{31}$

$a = 1103515245$

$c = 12345$

with effective C code

newran = [a * oldran + c] & MASK

where MASK has lower 31 bits = 1
all higher bits = 0.

we get a floating point number between 0 and 1 by:

floatran = 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

$1+1=0$
$1+0=0+1=1$
$0+0=0$

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 = seed for the congruential method

If you want a truely random start, a standard "dodge" is to set the seed to the binary 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 (binary, seed)
ranset (seed)
```

Correlations.

Often poorly chosen random number generators exhibit correlations between neighbouring random numbers

This is well illustrated by tests by Chin of integer generator supplied by Borland with Turbo Pascal v. 3-0

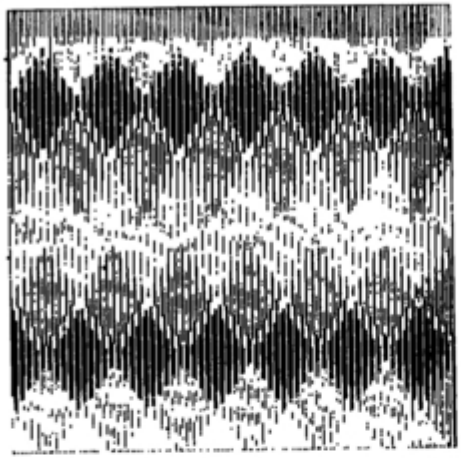


Fig. 2. More than 1000000 random numbers generated by Turbo-Pascal v3.0 and plotted as (x, y) pairs.

Adjacent random numbers considered as (x, y) pairs

Other generators are better!

Numerical Recipes has a program that removes any such problems

Essentially set up a list $R(i)$ of L random numbers.

Then choose j $\quad 1 \le j \le L$

Set 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 $\sigma^2 = 1/12$ and u_i are uniformly distributed between 0 and 1, is gaussianly distributed for large N. This for $N \sim 10$ 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

$$\exp\left[-\tfrac{1}{2}(g_1^2 + g_2^2)\right] dg_1\, dg_2 .$$

$g_1 = r\cos\theta$
$g_2 = r\sin\theta$

$$\exp\left(-\tfrac{1}{2}[g_1^2 + g_2^2]\right) dg_1\, dg_2$$

$$= \exp\left(-\tfrac{1}{2} r^2\right) r\, dr\, d\theta.$$

$$= \exp\left(-\tfrac{1}{2} r^2\right) d\left(\tfrac{1}{2} r^2\right) d\theta.$$

$$= d\left(\exp\left(-\tfrac{1}{2} r^2\right)\right) d\theta$$

$\therefore \quad u_1 = \exp\left(-\tfrac{1}{2} r^2\right) \qquad u_2 = \theta$

are uniform

$\Rightarrow g_1, g_2$ Gaussian.

Calculations for General Distributions

i) Example: exponential

Suppose x has distribution e^{-x} for $x = 0$ to ∞.

$$e^{-x}\,dx = d(e^{-x})$$

$y = e^{-x}$ is uniform.

$\therefore$ $x = -\ln y$ where y is a uniform random variable 0 to 1

ii) 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 here but if as above $p(x) = 0$ $x \leq 0$.

You must be able to invert integral to be able to do this

For $p(x) = e^{-x}$

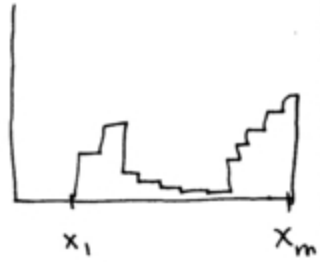
$w = 1 - e^{-x}$ is essentially as above because $e^{-x} = 1 - w$ as same as $e^{-x} = y$ as y and $1-w$ are both

uniform between 0 and 1.

iii) Interpolation:

Suppose $p(x)$ is given by a table

x_1	$p(x_1)$
x_2	$p(x_2)$
$\vdots$	
x_m	$p(x_m)$



in say equally spaced intervals
$x_k - x_{k-1} = \delta x$.

Then you form cumulative distribution

$x_1 + \delta x$	$p(x_1)$	$F(x_1 + \delta x)$
$x_1 + 2\delta x$	$p(x_1) + p(x_2)$	
$\vdots$		
x_{m+1}	$\sum_{i=1}^{m} p(x_i)$	$= 1$ if normalized

Then generate random number r
$0 \le r \le 1$.

Then use inverse interpolation to find x

$$F(x) = r$$

Then x is distributed according to $p(x)$

(iv) Accept / Reject

Then $p_{max} = \max_x p(x) \qquad x_1 \le x \le x_2$

Let w_{max} be any number $> p_{max}$

Generate x uniformly in $[x_1, x_2]$.

Generate r uniformly in $[0\ 1]$

If $r > \dfrac{p(x)}{w_{max}}$, ignore x and start again

$\le \dfrac{p(x)}{w_{max}}$, accept x.

(v) One can combine this with technique in (i).

Suppose $f(x)\,dx = d(F(x))$

and we can invert $F(x)$ as in (i)

Let w_{max} be any number $\geq \max_x p(x)/f(x)$

Then generate x in $[x_1, x_2]$ distributed according to $f(x)$ using technique in (i)

Use (iii) with $p(x) \to p(x)/f(x)$

to "correct" $f(x)$ to $p(x)$

Clearly one chooses $f(x)$ to have

a distribution "near" ~~as~~ to $p(x)$ but with a simple enough form that one can find and invert it

(vi) 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 the production probability. Then if if we want to calculate some simple quantity q_i, the best is to take

$$\langle Q \rangle = \frac{\sum_{i=1}^{N} q_i w_i}{\sum_{i=1}^{N} w_i} \qquad (1)$$

An alternative is to use acceptance-rejection to find a set M of events j with "weight 1".

$$M < N$$

Then

$$\langle Q \rangle = \frac{1}{M} \sum_{j=1}^{M} q_j \qquad (2)$$

When is (2) superior to (1).

If you are just calculating $\langle Q \rangle$, then (1) ~~usually~~ contains 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. $0.001\, w_{max}$!

Maximum Likelihood is <u>Best</u>! (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 λ characterising the distribution. We can assume that $\langle S \rangle \sim \lambda$ (if $\langle S \rangle \sim \lambda^3$ then replace S by $\sqrt[3]{S}$ etc.)

The <u>bias</u>

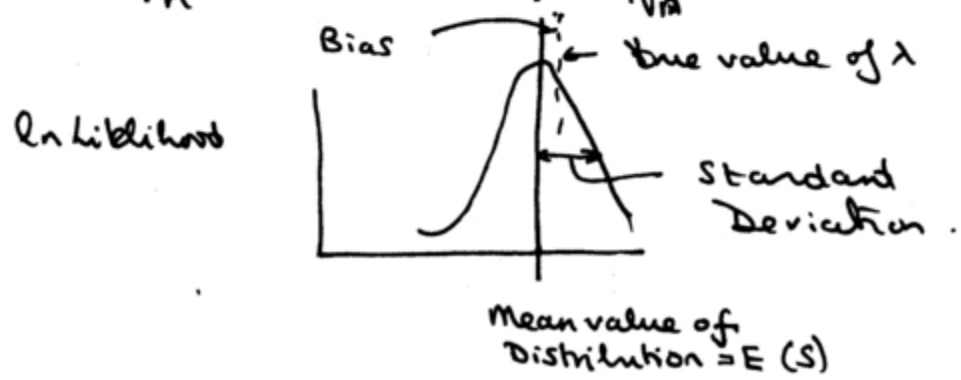
$$B(\lambda) = E(S) - \lambda$$

This should be distinguished from the variance $\sigma^2(S)$

$$\sigma^2(S) = \langle (S - E(S))^2 \rangle$$

For the maximum likelihood method

$B(\lambda) \sim 1/n$ and $\sigma(S) \sim 1/\sqrt{n}$



As the statistical error is $\propto \sqrt{n}$ times bigger than the bias, one usually ignores it. The ~~small~~ maximum likelihood method has asymptotically small bias and the minimum possible variance. We can only bound 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 wrt λ, we get

$$1 + B'(\lambda) = \int S \left[\partial/\partial\lambda (\ln L)\right] \exp(\ln L)\, dx_1 \ldots dx_n$$

$$= \left\langle S \left[\partial/\partial\lambda (\ln L)\right]\right\rangle$$

as $\left\langle \partial/\partial\lambda (\ln L)\right\rangle = \partial/\partial\lambda \int \exp(\ln L)\, dx_1 \ldots dx_n$

$$= \partial/\partial\lambda \; 1$$

by normalization of probabilities

$$= 0$$

we can replace S by S - any constant

or in particular $S - \langle S \rangle$

$$\therefore \quad 1 + B'(\lambda) = \left\langle \left[(S - \langle S \rangle) \, \partial/\partial\lambda (\ln L) \right] \right\rangle \qquad (\P)$$

Now for any random variables α, β

we have $\langle \alpha\beta \rangle^2 \leq \langle \alpha^2 \rangle \langle \beta^2 \rangle \qquad (*)$

This is ~~of course~~ obvious because for

any constants a, b

$$\langle (a\alpha + b\beta)^2 \rangle = a^2 \langle \alpha^2 \rangle + 2ab \langle \alpha\beta \rangle + b \langle \beta^2 \rangle \geq 0$$

This is implies $(*)$.

Applying this inequality to $(\P)$

we have

$$[1 + B'(\lambda)]^2 \leq \sigma^2(S) \left\langle \left[\partial/\partial\lambda (\ln L) \right]^2 \right\rangle$$

$$\text{or} \quad \sigma(S) \geq \frac{1 + B'(\lambda)}{\sqrt{\left\langle \left[\partial/\partial\lambda (\ln L) \right]^2 \right\rangle}} \qquad \text{(CR)}$$

We set the denominator as $\sqrt{I(\lambda)}$
where the "information" $I(\lambda)$ is given by:

$$I(\lambda) = \int \left[\sum_{i=1}^{n} \partial/\partial\lambda (\ln P(x_i, \lambda))\right] \cdot \left[\sum_{j=1}^{n} \partial/\partial\lambda (\ln P(x_j, \lambda))\right] L(x_1, x_2 \ldots x_n) \, dx_1 \ldots dx_n$$

The off diagonal terms are just

$$\langle \partial/\partial\lambda (\ln P(x_i, \lambda)) \rangle \langle \partial/\partial\lambda (\ln P(x_j, \lambda)) \rangle$$

$= 0$ by same argument as on the bottom of ML2.

$$\therefore I(\lambda) = n \underbrace{\int \left[\partial/\partial\lambda (\ln P(x, \lambda))\right]^2 P(x, \lambda) \, dx}$$

This can also be written as

$$-\int \frac{\partial^2}{\partial\lambda^2} (\ln P(x, \lambda)) \, P(x, \lambda) \, dx$$

$I(\lambda)$ is information on λ contained in 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 Cramer Rao bound) states that for an unbiased statistic

$$\sigma(S) \geq 1/\sqrt{I(\lambda)}.$$

Relevance to Maximum Liklihood Method

The maximum liklihood statistic S is given by the equation

$$\frac{\partial}{\partial S} \ln L(S, x_1, x_2 .. x_n) = 0$$

or

$$\sum_{i=1}^{n} \frac{\partial}{\partial S} \ln P(x_i, S) = 0. \qquad (1)$$

Expand $\ln P(x_i, S)$ about $S = \lambda$

$$\ln P(x_i, S) = \ln P(x_i, \lambda) + (S-\lambda) \frac{\partial}{\partial \lambda} \ln P(x_i, \lambda) \qquad (2)$$

$$+ \tfrac{1}{2}(S-\lambda)^2 \frac{\partial^2}{\partial \lambda^2} \ln P(x_i, \lambda) + \ldots$$

we will return to this

Substitute (2) into (1), to find:

$$S - \lambda = \frac{\sum_{i=1}^{n} \partial/\partial\lambda \ln P(x_i, \lambda)}{\sum_{i=1}^{n} -\partial^2/\partial\lambda^2 \ln P(x_i, \lambda)}$$

This is of our standard form

$$\frac{\sum_{i=1}^{n} f(x_i)}{\sum_{i=1}^{n} g(x_i)} \qquad f = \partial/\partial\lambda \ln P(x, \lambda) \qquad g = -\partial^2/\partial\lambda^2 \ln P(x, \lambda)$$

$$\langle S - \lambda \rangle = \langle f \rangle / \langle g \rangle = 0 \quad (\langle f \rangle = 0 \text{ by } (*)$$

argument on bottom of ML2).

$$\langle (S-\lambda)^2 \rangle = \frac{1}{n\langle g \rangle^2} \left\langle \left(f - \frac{\langle f \rangle g}{\langle g \rangle} \right)^2 \right\rangle$$

$$= \frac{\langle f^2 \rangle}{n \langle g \rangle^2}$$

but $\langle f^2 \rangle = \langle g \rangle = \frac{I(\lambda)}{n}$ (See middle of ML4).

$$\therefore \quad \langle (S - \lambda)^2 \rangle = \frac{1}{I(\lambda)}. \qquad (**)$$

i.e. the maximum likelihood method is
Unbiased — See (*)
Optimal i.e. Variance = Cramer Rao bound (**).

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 $\langle S-\lambda\rangle = 0$ derived on the previous page. The first is the truncation involved in (2). In fact:

$$\left[-\sum_{i=1}^{n} \partial^2/\partial\lambda^2 \ln P(x_i,\lambda)\right](S-\lambda)$$

$$= \sum_{i=1}^{n} \partial/\partial\lambda \ln P(x_i,\lambda) + \underbrace{\tfrac{1}{2}(S-\lambda)^2 \sum_{i=1}^{n} \partial^3/\partial\lambda^3 \ln P(x_i,\lambda)}$$

Here we can replace $(S-\lambda)^2$ by $1/I(\lambda)$ and $\sum_{i=1}^{n} \partial^3/\partial\lambda^3 \ln P(x_i,\lambda)$ by $n\left\langle \partial^3/\partial\lambda^3 \ln P(x,\lambda)\right\rangle$ with an error of $O(1/\sqrt{n})$ as ~~the~~ each of random variables $(S-\lambda)^2$ and $\sum_{i=1}^{n} \partial^3/\partial\lambda^3 \ln P(x_i,\lambda)$ is peaked about its mean with error of $O(1/\sqrt{n})$.

e.g. in our standard example

$$Fg = (n\langle f\rangle + \sqrt{n}\,\tilde{f})(n\langle g\rangle + \sqrt{n}\,\tilde{g})$$

with $\langle\tilde{f}\rangle = \langle\tilde{g}\rangle = 0$.

$\therefore Fg = n^2\langle f\rangle\langle g\rangle + O(n\sqrt{n})$ term with zero mean and $O(n)$ term with nonzero mean

Consider F/G with $\langle f \rangle = 0$

$$\langle F/G \rangle = \left\langle \left\{ \frac{\sum_{i=1}^{n} f(x_i)}{n\langle g \rangle + \sum_{i=1}^{n} [g(x_i) - \langle g \rangle]} \right\} \right\rangle$$

$$\sim \frac{1}{n\langle g \rangle} \left\langle \sum_{i=1}^{n} f(x_i) \left[1 - \frac{1}{n\langle g \rangle} \sum_{i=1}^{n} [g(x_i) - \langle g \rangle] + \cdots \right] \right\rangle$$

$$= \frac{-1}{n\langle g \rangle \, n\langle g \rangle} \left\langle \sum_{i=1}^{n} f(x_i) \sum_{j=1}^{n} [g(x_j) - \langle g \rangle] \right\rangle$$

irrelevant as multiplies term with zero mean

only $i = j$ counts as x_i independent

$$= \frac{-1}{n^2 \langle g \rangle^2} \left\langle \sum_{i=1}^{n} f(x_i) g(x_i) \right\rangle$$

$$= -\frac{1}{n \langle g \rangle^2} \langle fg \rangle \qquad \Leftarrow$$

i.e. when f and g are correlated ~~with~~ we get a $O(1/n)$ correction to mean.

Apply this ~~to~~ with $f = \partial/\partial\lambda \ln P(x, \lambda)$

$$g = -\partial^2/\partial\lambda^2 \ln P(x, \lambda).$$

We finally get

$$B(\lambda) = \langle (S-\lambda) \rangle = + \frac{n}{2I^2(\lambda)} \left\langle \frac{\partial^3}{\partial\lambda^3} \ln P(x,\lambda) \right\rangle \qquad (*)$$

$$+ \frac{n}{I^2(\lambda)} \left\langle \frac{\partial}{\partial\lambda} \ln P(x,\lambda) \frac{\partial^2}{\partial\lambda^2} \ln P(x,\lambda) \right\rangle$$

as $I(\lambda) \propto n$, we see that $B(\lambda) \sim 1/n$.

The above can be written in a different form using

$$\frac{\partial}{\partial\lambda} \left\langle \frac{\partial^2}{\partial\lambda^2} \ln P(x,\lambda) \right\rangle = \frac{\partial}{\partial\lambda} \int \frac{\partial^2 \ln P(x,\lambda)}{\partial\lambda^2} P(x,\lambda)\, d\mathbf{X}$$

$$= \left\langle \frac{\partial^3}{\partial\lambda^3} \ln P(x,\lambda) \right\rangle + \left\langle \frac{\partial}{\partial\lambda} \ln P(x,\lambda) \frac{\partial^2}{\partial\lambda^2} \ln P(x,\lambda) \right\rangle$$

but I think (*) is more convenient.

Note that all the terms in (*) <u>can</u> be estimated from a given data sample and so bias can be corrected for. As $B(\lambda) \sim 1/n$ and estimate has statistical error $1/\sqrt{n}$, one can in fact find $\hat{\lambda}$ with a bi

with a bias $\sim 1/n\sqrt{n}$ in a given experiment. Normally one does not do this because statistical error is of $O(1/\sqrt{n})$ and outweighs bias.

<u>example</u>:

if $P(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp[-(x-\mu)^2/2\sigma^2]$

where x is observed and we wish to find μ, σ then maximum likelihood estimates are

$$S(\mu) = \sum_{i=1}^{n} x_i / n$$

$$S(\sigma^2) = \sum_{i=1}^{n} (x_i - S(\mu))^2 / n$$

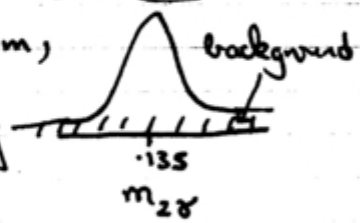
but in fact we know that $S(\sigma^2)$ is biased and unbiased formula is $n/(n-1)$ times $S(\sigma^2)$. This correction is $O(1/n)$ by above theorem and indeed can be calculated from it. (after we do trivial extension to ~~several~~ a vector of theoretical parameters $\underline{\lambda}$).

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 Jerry Friedman at a CERN Computer School.

Non parametric Data Representation

Given a mass spectrum,



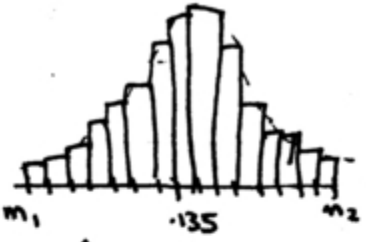
then a <u>parametric</u> summary of this is to assume that the number of counts (probability) as a function of mass can be fitted

to a Gaussian plus a polynomial background.

$$p(m) = c \exp\left[-(m-m_0)^2/2\sigma^2\right] + \sum_{i=0}^{N} a_i (m-m_0)^i$$

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.



i.e. divide mass range into equal intervals of size Δm and count the number n_i of counts in the i'th bin $m_1 + (i-1)\Delta m \to m_1 + i\Delta m$ for $i = 1 \ldots (m_2 - m_1)/\Delta m$. 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 D4/40). in a way that is clumsier but free of ad hoc (and possibly wrong) assumptions about the functional form of $p(m)$.

Let us now approach the problem more formally: we have measurements $\{x_i \; i = 1 \ldots N\}$ where the x_i are independent random

variables distributed according to some <u>unknown</u> probability density function $p(x)$. We wish to form estimates $\hat{p}(x) = \varphi_N(x_1 \ldots x_N)$ of $p(x)$.

§1: Histogram Method

Define $g_i(x)$ to be 1 in ith bin and zero elsewhere ($g_i(x)$ is called an indicator function).
i.e. if range $a \to b$ and $\Delta x = (b-a)/m$, then

$$g_i(x) = 1 \qquad a + (i-1)\Delta x \leq x \leq a + i.\Delta x$$
$$ = 0 \qquad \text{elsewhere}$$

Then we define our estimate of $p(x)$ as

$$\hat{p}_{Histogram}(x) = \sum_{i=1}^{m} \frac{1}{N}\left(\sum_{j=1}^{N} g_i(x_j)\right) g_i(x)$$

In each bin; $\hat{p}_H$ will for large N tend to its mean = $\int_{bin} p(x)\,dx$ with as standard deviation that this mean ÷ $\sqrt{\text{number of } x_i \text{ in bin}}$

$$\sim \frac{\sqrt{\int_{bin} p(x)\,dx}}{\sqrt{N}}.$$

(See argument on D33 1/4 for a more accurate analysis based on correct binomial distribution).

This method has the following disadvantages

(1) As $N \to \infty$, $\hat{p}_H(x)$ does not $\to p(x)$ but rather to $\sum_{i=1}^{m} g_i(x) \int_{i\text{-th bin}} p(x)\,dx$

This can only be cured by letting bin size $\to 0$ as $N \to \infty$.

(2) As bin size uniform some bins

will have lots of entries (where $p(x)$ big) and others a small number. One would like to automatically choose larger bin size where $p(x)$ small.

§2 Orthogonal Functions

A generalization of §1 is to replace the indicator functions $g_i(x)$ by general orthogonal functions $\psi_i(x)$

$$\int \psi_i(x)\,\psi_j(x)\,dx = \delta_{ij}$$

$$\hat{p}_0(x) = \sum_{i=1}^{m} \left[\int \psi_i(x)\,p(x)\,dx\right] \psi_i(x)$$

$$= \frac{1}{N} \sum_{i=1}^{m} \left(\sum_j \psi_i(x_j)\right) \psi_i(x).$$

Clearly the histogram formulae are special cases of this where

$\psi_i(x)$ are just indicator functions. By the central limit $\hat{p}_0(x)$ has a nice limit as $N \to \infty$ $\left(\frac{1}{N} \sum_j \psi_i(x_j) \to \int \psi_i(x) p(x)\, dx \right.$ with error $\propto \left. \frac{1}{\sqrt{N}} \right)$ although the coefficients of $\psi_i(x)$ are no longer statistically independent. Again $\hat{p}_0(x)$ does <u>not</u> $\to p(x)$ unless you take $M \to \infty$ as $N \to \infty$ so that all terms are included in $\psi_i(x)$ expansion of $p(x)$.

§3 Empirical Cumulative Distribution

Define

$$\hat{F}(x) = \begin{cases} 0 & x < x_1 \\ i/N & x_i \le x \le x_{i+1} \\ 1 & x \geq x_n \end{cases}$$

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 $\to F(x)$ as $N \to \infty$.

This empirical distribution will be very useful in goodness of fit tests. Now we use it to find another estimate of $p(x)$

§4 The Rosenblatt Estimator

We just define $\hat{p}_R(y)$ as the naive derivative of $\hat{F}(y)$

$$\hat{p}_R(y) = \frac{1}{2h}\left[\hat{F}(y+h) - \hat{F}(y-h)\right] \quad (4.1)$$

Actually this is just number of points in bin $y-h \le x \le y+h$ 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 neighbouring bins of a histogram where

uncorrelated.

One can find the mean and variance of $\hat{p}_R(y)$ as

$$\langle \hat{p}_R(y) \rangle = \frac{1}{2h}\left[F(x+h) - F(x-h)\right] \quad (4.2)$$

$$= p(y) + \frac{h^2}{6} p''(y) + \cdots \quad \text{(P.T.O.)}$$

$$\langle (\hat{p}_R(y) - p(y))^2 \rangle = \underset{①}{\frac{p(y)}{2hN}} + \underset{②}{\frac{h^4}{36} |p''(y)|^2} \quad (4.3)$$

These ~~estimate~~ results are errors on using

$$\hat{p}_R(y) = \frac{1}{2hN} \sum_{i=1}^{N} \psi_0\left(\frac{y - x_i}{h}\right)$$

where $\psi_0(x) = 1$ for $-1 \le x \le 1$
$= 0$ otherwise.

① is just standard $\sigma \propto \sqrt{\#\text{events}}$ 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 had

calculated behaviour of $\hat{p}_R(y)$ wrt lin integral $\frac{1}{2h}[F(x+h)-F(x-h)]$: we would have found similar terms for the histogram method if we had ~~calculated~~ viewed it as an estimate of $p(y)$ and not the lin integral.

From (4.3) we want $h \to 0$ while $hN \to \infty$. We can minimize (4.3) for fixed $p(y)$, $p''(y)$ by choosing

$$h = C N^{\alpha} \qquad (4.4)$$

Putting this into right hand side of (4.3) and requiring $\partial/\partial C$ and $\partial/\partial\alpha = 0$, we find $\alpha = -1/5$ (so that both terms on rhs go like $N^{-4/5}$) and

$$C = \left[9p(y)/2\,|p''(y)|^2 \right]^{1/5} \qquad (4.5)$$

G5 Parzen Estimates

These are just like the Rosenblatt case except that $\mathcal{G}$ defined below (4.3) is replaced by any function whose integral is 2 from $-\infty$ to $+\infty$.

Possible choices are

$$\mathcal{G}_0(x) = \exp(-|x|)$$

$$\text{or} = \frac{2}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$$

These generalized estimates have very similar properties to Rosenblatt's.

§6 k-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 Rosenblatt's method. We choose h so that interval $y-h \le x \le y+h$ 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

$$\hat{p}_k(y) = k(N)/2Nh(y,k) \qquad (6.1)$$

where h is determined for each y as above.

Qualitatively this method is reasonable i.e. h gets large where p(x) is small and it overcomes one of the major problems with the histogram method.

One can show that

$$\langle \hat{p}_k(y) \rangle = p(y) + \frac{1}{24}\left[\frac{k(N)}{N}\right]^2 \frac{p''(y)}{p^2(y)} \qquad (6.2)$$

$$\langle (\hat{p}_k(y) - p(y))^2 \rangle = \frac{p^2(y)}{k(N)} + \left[\frac{p''(y)}{24\, p^2(y)} \left(\frac{k(N)}{N}\right)^2\right]^2 \qquad (6.3)$$

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 $k(N) \to \infty$ but
$\frac{k(N)}{N} \to 0$
and we can estimate best $k(N)$ just as for h to get

$$\left.\frac{k(N)}{N}\right|_{best} = \left\{ \frac{144\, p^6(x)}{|p''(x)|^2 N} \right\}^{1/5}$$

and $\frac{k(N)}{N}$ behaves in same 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 ϕ_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 $\hat{p}_k(y)$ at neighbouring 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.

§10 Liklihood Ratio Method

We wish to test if a ~~particular~~ theoretical parameter λ has a particular value $\lambda^{(0)}$. Now $H^{(1)}$ is all other values of λ. We find

T = Value of liklihood maximized over all values of λ ÷ Value of liklihood calculated for $\lambda = \lambda^{(0)}$.

Note that in this ratio all the dubious undefined parts of L

cancelled out.

The value of T is a generalization of χ^2 test ($-2\ln T = \chi^2_{\text{minimized}} - \chi^2$ calculated for $\lambda = \lambda^{(0)}$) ~~but~~ and clearly $T \sim 1$ ~~is a success~~ 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 L becomes Gaussian in λ). Brandt does the calculation for a Gaussian distribution and derives the Student's t distribution

from this more general point of view.

§11 Methods based on Cumulative Distribution

These avoid "binning" necessary in χ^2 method but are only [redacted]ble in 1 dimension. We start [redacted]th the empirical cumulative distribution

$$\hat{F}(x) = \frac{1}{n}\sum_{j=1}^{n} \Theta(x_j - x)$$

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

$$\omega^2 = \int_{-\infty}^{+\infty} \left[\hat{F}(x) - F(x)\right]^2 p(x)\,dx$$

Rather remarkably, the distribution of ω^2 is <u>independent</u> of $p(x)$* and one can show that

$$\langle \omega^2 \rangle = 1/6n$$

$$\sigma^2_{\omega^2} = \frac{4n-3}{180n^3}$$

* This is not intuitively obvious to me but can be shown because
1) all moments of $\hat{F}(x)$ only depend on $F(x)$
2) then we can let $z = F(x)$ in integral to remove F dependence (note $dz = p(x)dx$).

One can also use

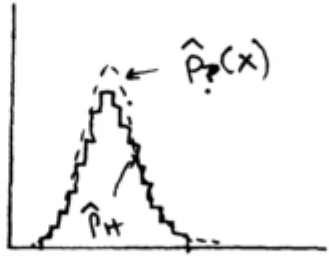
$$D = \max_{-\infty \le x \le \infty} |F(x) - \hat{F}(x)|$$

~~For large n, D~~ but and this is discussed on p. 270 of EDJRS. Again the distribution of D is independent of $F(x)$.

§12 Rootograms

In a normal histogram, where the i'th bin contains N_i events, the error is proportional to $\sqrt{N_i}$. The variation of the size of error with population makes it hard to judge by eye the goodness of fit of some estimate $\hat{p}_?(x)$ to the histogram $\hat{p}_H(x)$

↑ maybe from maximum liklihood



One way around this is to plot $\alpha_i = \sqrt{N_i}$ for

$$\delta\alpha_i = \sqrt{N_i + \delta N_i} - \sqrt{N_i}$$

$$\sim \frac{\delta N_i}{\sqrt{N_i}} = 1.$$

So a plot of the α_i 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 rootogram. Friedman suggests plotting the residuals

$$r(x) = \sqrt{\hat{p}_?(x)} - \sqrt{\hat{p}_H(x)}$$

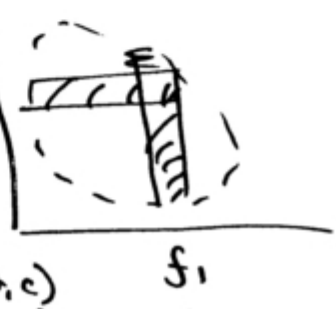
as being even clearer. [$r(x)$ should have mean zero and an error independent of x i.e. if $\hat{p}_H(x) = c\,N(x)$ then standard deviation of $r(x)$ is $\sqrt{c}$].

We found on one experiment where we were plotting electromagnetic shower pulse heights (& fitting them to predetermined forms) that ~~we got a greater dy~~ c the dynamic range $\sqrt{p_{min}} \to \sqrt{p_{max}}$ ~~we~~ was visually far ~~easier~~ better than $p_{min} \to p_{max}$ (e.g. in our case $p_{min} \approx 10$ cts $p_{max} \approx 1000$ cts and so a linear histogram had a dynamic range ~ 100 whereas the rootogram only varied by a factor of 10).

§13 Multivariate ~~the~~ mapping methods

The previous techniques for non parametric representation do not easily generalize to ~~multip~~ more than 1 dimension i.e. when the observable x_i is a vector in d dimensions.

Typically one tackles vector valued observables by forming functions $y = f(\underline{x})$ and using univariate techniques on y. Most histogramming packages also allow one to examine correlations (using a "Scatterplot") between two such f's. i.e. $y = f_1(\underline{x})$ and $z = f_2(\underline{x})$



if $\underline{x}$ corresponds to particle parameters (p_a, p_b, p_c for 3 particles a, b, c) in a 3 particle final state

then perhaps

$$y \text{ is } (p_a + p_b)^2$$
$$z \text{ is } (p_b + p_c)^2$$

and scatterplot is usually called the Dalitz plot.

One can choose the function $f(\underline{x})$ 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 $\underline{x}_i$. We will only discuss the latter methods here.

A) Principal Components

Here one chooses the general linear form $y = \hat{\underline{a}} \cdot \underline{x}$. For each $\underline{a}$, we calculate the standard deviation of the univariate distribution of y.

if $V_{kl} = \frac{1}{N}\sum_{i=1}^{N} x_k^{(i)} x_l^{(i)} - \left[\frac{1}{N}\sum_{i=1}^{N} x_k^{(i)}\right]\left[\frac{1}{N}\sum_{i=1}^{N} x_l^{(i)}\right]$

is the usual co-variance matrix of the observations $\underline{x}^{(i)}$, then.

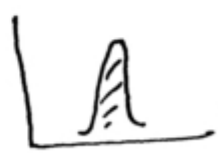
$$\sigma(y) = \sqrt{\underline{a}^T \underline{\underline{V}}\, \underline{a}} \qquad (*)$$

The principal axis method gives the projection axis $\underline{a}$ that maximizes $\sigma(y)$. This clearly corresponds to the eigenvector of largest eigenvalue for $\underline{\underline{V}}$. Continuing this process we deduce that the natural axes for examining our data are just the eigenvectors of the correlation matrix $\underline{\underline{V}}$.

e.g. y distributions vary from

$y = a_{max} \cdot \underline{x}$

. . . . . .



$y = \underline{a}_{min} \cdot \underline{x}$

largest standard deviation $\longrightarrow$ 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).

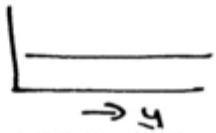
B: Projection Pursuit. (Friedman and Tukey, IEEE transactions on Computers C-23, 881 (74))

The idea here is that we wish to look for seperated clusters and the above principal axis method emphasized seperation but not clustering. So we try to maximize

$$I(\hat{a}) = \sigma(\hat{a})\, d(\hat{a})$$

where $\sigma(\hat{a})$ is given by (*) on G37. $d(\hat{a})$ is to be chosen to be larger in situations like

than for



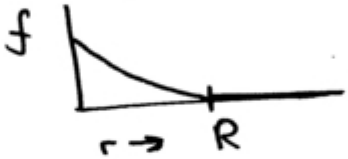
Friedman and Tukey choose

$$d(\hat{a}) = \sum_{i,j=1}^{N} f(r_{ij})$$

where r_{ij} is distance from $\underline{x}_i$ to $\underline{x}_j$ projected on axis $\hat{a}$

$$r_{ij}^2 = |\hat{a}.(\underline{x}_i - \underline{x}_j)|^2$$

and f is any smooth function of the form



$f = 0 \quad r > R$

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 $I(\hat{a})$ ~~is a sh~~ can be performed by standard non linear maximization methods – it cannot be solved analytically as for choice of $\sigma(\hat{a})$ as quantity to maximize.

The method can be generalized to more than one direction $\hat{a}$. Thus one replaces $I(\hat{a})$ by

$$I(\hat{\underline{a}}, \hat{\underline{b}}) = \sigma(\hat{a})\,\sigma(\hat{b})\,d(\hat{a}, \hat{b})$$

where in d one calculates

$$r_{ij}^2 = |\hat{a}\cdot(\underline{x}_i - \underline{x}_j)|^2 + |\hat{b}\cdot(\underline{x}_i - \underline{x}_j)|^2$$

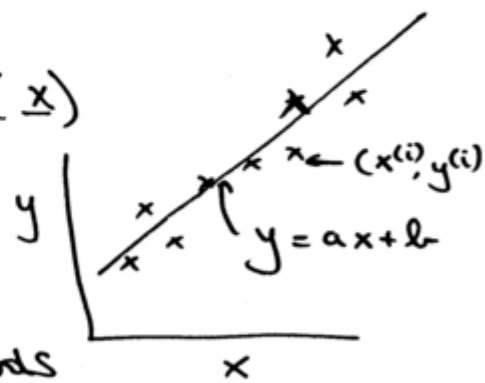
and take $\hat{a}, \hat{b}$ to be orthogonal directions.

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 all the tools necessary for this yet (we need to discuss criterion for comparing

two ~~two~~ samples from ~~two~~ possibly different multivariate distributions. However we can discuss the slightly simpler problem of "smoothing" i.e. ~~given~~ we ~~a~~ suppose that a "response" $y^{(i)}$ is associated with each of our random variables $\underline{x}^{(i)}$. We wish to find an estimate of the supposed functional relation:

$$y = \varphi(\underline{x})$$

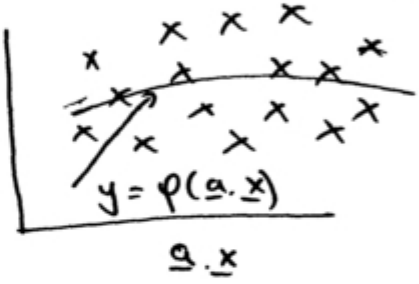


In the 1D case, there are various well known ~~forms~~ methods (least squares fit ≈ regression analysis, splines) which we will hopefully discuss. To generalize to the multidimensional

case, we suppose that

$$y = \varphi(\underline{a}\cdot\underline{x})$$

For each $\underline{a}$, we apply the univariate smoothing algorithm to the points $(\underline{a}\cdot\underline{x}^{(i)}, y^{(i)})$. In general this will <u>not</u> look very good as y is multivalued as a function of $\underline{a}\cdot\underline{x}$



So we choose $\underline{a}$ to minimize this: i.e. to minimize scatter about the smooth i.e. $\underline{a}$ is vector that minimizes

$$\sum_{i=1}^{n} \left(y^{(i)} - \varphi(\underline{a}\cdot\underline{x}^{(i)})\right)^2 \qquad (*)$$

Again this needs a nonlinear $\begin{bmatrix}\text{min}\\ \text{max}\end{bmatrix}$inizer.

Calling this vector $\underline{a}^{(1)}$ we replace

$$y^{(i)} \to y^{(i)} - \varphi_{min}(\underline{a}^{(1)}\cdot\underline{x}^{(i)})$$

and repeat procedure until the residual sum $(*)$ is small.

§14 Minimal Spanning Trees: (MST)

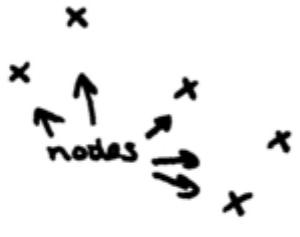
See J. L. Bentley and J. H. Friedman
Ieee Trans on Comp C-27, 97 (78)
J. Dorfan, "A Cluster Algorithm for the Study of Jets in High Energy Physics", SLAC - 2623 (1980).

Given a bunch of points in a multidimensional space on which a metric has been defined i.e. d_{ij} is distance between $\underline{x}^{(i)}$ and $\underline{x}^{(j)}$. Then the MST is gotten by joining points $\underline{x}^{(i)}$ (called nodes) by edges in such a way that

(i) There is a connected path between any two nodes

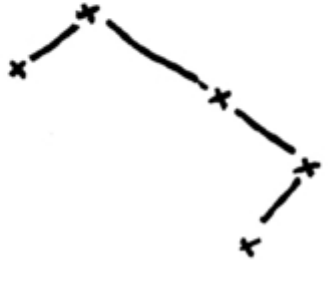
(ii) The sum of the lengths of the edges is the smallest possible

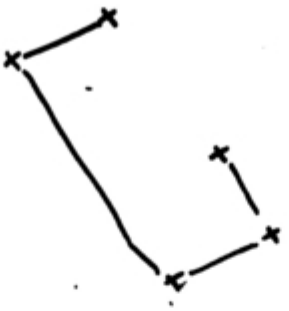
e.g in



with distance as euclidean metric

then MST is:

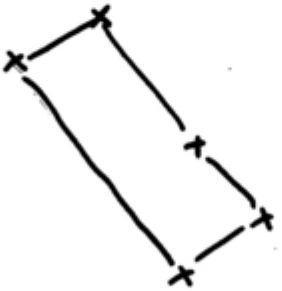


and 

is not an MST as the Σ lengths of connected edges is longer than optimal solution.

and 

is not an MST as there is no path to A

and 

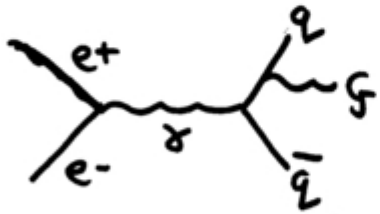
is not an MST as there is an unnecessary edge.

It can be shown than an MST always exists and is essentially unique (i.e. only if some of the d_{ij} are equal can there be more than one MST e.g.

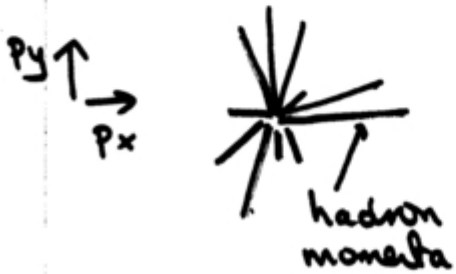
if $d_{CE} = d_{DE}$

then one can complete MST by joining either $C \to E$ or $D \to E$.

MST's have many uses that don't 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 (71)) discusses use of MSTs to find general patterns. Dorfan (See D43) uses the MST to find jets in e^+e^- collisions



appears as a bunch of hadrons loosely clustered around original parton directions. If we establish a metric $d_{ij} = |p_\perp|$ of particle i wrt j (where i is that one of i,j which is lower momentum and j is higher momentum particle).



(Dorfan uses a different metric: ours is motivated by the limited $p_\perp$ of particles wrt original jet axis). 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 ~~from~~ decompose into jets. The same technique is used by Zahn to break a general bunch 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 Bentley and Friedman show ~~that~~ MSTs can be found in the typical NlogN sorting time.

G1[redacted]5 Comparison of Multivariate Distributions

We now return to the problem discussed in G10,11 for the univariate case.

i.e. This problem has two variants.

(1) Given two sets of points $\{\underline{x}^{(1,i)} : i=1..N_1\}$ $\{\underline{x}^{(2,j)} : j=1..N_2\}$, are these two sets drawn from the same probability distribution?

(2) Given one set of points $\{\underline{x}^{(i)} : i=1..N\}$ - is it drawn from a given probability distribution $p(\underline{x})$? This second case can be reduced to the first if one uses a Monte Carlo method to generate a set of points according to $p(\underline{x})$. In this case $p(\underline{x})$ could be the probability distribution found from the maximum likelihood method

A: k'th Nearest Neighbour Comparison of two Sets

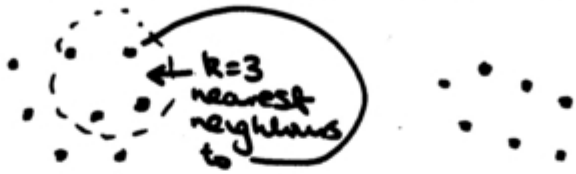
Join our two sets into a sample of size $N_1 + N_2$ but remember from which of two sets each point comes from. Fix k.

Now look at each point and find the k points from joint sample that are nearest to the point. Record the number $n_1(k)$ of points from set 1 that are among our k nearest neighbours.

Consider the N_1 values $S_1: \{n_1(k)\}$ of set 1 counts among kth nearest neighbours of set 1 points: and N_2 values $S_2: \{n_2(k)\}$ of set 1 counts among kth nearest neighbours of set 2 points. Then if the two sets vector $\{x^{(i,j)}\}$ are really the same, the two integer sets S_1, S_2 are also the same. The point is that we here swopped multivariate comparison to a univariate comparison. The latter can be performed by simple modifications of the methods described in §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 $\{\underline{x}^{(1,i)}\}$ bear no resemblance to each other



Set 1: Set 2:

then nearest neighbours to a Set 1 point are all Set 1 and to a Set 2, all Set 2.

Then S_1 is $(k, k \ldots k)$
S_2 is $(0, 0 \ldots 0)$

~~and in fact they agree even though the original vector sets were distinct.~~

and the integer sets are very dissimilar.

If the vector sets are in fact from the same distribution, then the distribution of $r_1(k)$ is - near both

a by set 1 and set 2 point the binomial corresponding to k trials with probability $N_1/(N_1+N_2)$ of success. This is not quite right because the set 1 points in different neighbourhoods overlap; one would have to do a Monte Carlo calculation to find the distribution of $n_1(k)$ for a particular multivariate distribution $p(\underline{x})$. The binomial form is usually good and one can compare S_1: $\{n_1(k)$ for set 1 centers$\}$ directly with this distribution using standard univariate methods for comparing a sample and an hypothesized distribution (§21).

B: k'th Nearest Neighbour Comparison between a Sample Set and a Supposed distribution $p(\underline{x})$ (problem 2 on p.§47)

~~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

P.T.O.

$V^{(i)}$ = volume of k'th nearest neighbour set (taken as a sphere) about point (i)

and $$v^{(i)} = \int_{\substack{k\ nn.\\ \text{volume}\\ \text{about } \underline{x}^{(i)}}} \frac{1}{p(\underline{x}')}\, p_{actual}(\underline{x}')\, d\underline{x}'$$

$$= \frac{1}{N} \sum_{j=1}^{N} \frac{1}{p(\underline{x}^{(j)})} \quad \begin{matrix} 1 \text{ point in volume} \\ 0 \text{ point outside volume} \end{matrix}$$

($p(\underline{x})$ is hypothesised distribution: p_{actual} is true distribution)

$$= \frac{1}{N} \sum_{\substack{k+1 \text{ points incl. center}\\ \text{inside nn volume}}} \frac{1}{p(\underline{x}^{(j)})}.$$

* More precise seems ~~average of~~ sphere whose volume is average of k and k+1 th nearest neighbour spheres.

Now $\{V^{(i)}\}$ $\{V^{(i)}\}$ can be compared by our univariate methods again.

This method only tests p near our sample points $\underline{x}^{(i)}$; if p is very large in regions where $x^{(i)}$'s do not exist, then one will only indirectly see this through the normalization of p. This defect is <u>not</u> present in the alternate method of generating points according to $p(\underline{x})$ and apply the two sample comparison methods.

Note this nn methods is 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, $1 \le i \le N_1 + N_2$ set

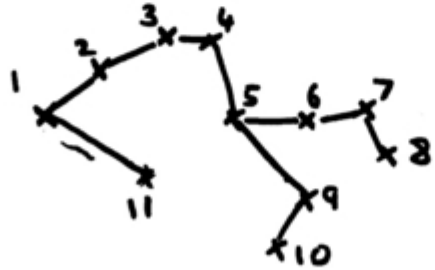
$$d_i = \frac{\left(\begin{array}{c}\#\text{ Set 1 points whose} \\ \text{position in list is} \le i\end{array}\right)}{N_1} - \frac{\left(\begin{array}{c}\#\text{ Set 2 points whose} \\ \text{position in list is} \le i\end{array}\right)}{N_2}$$

$$D = \max_i |d_i|.$$

We generalize this to the multivariate case by taking our vectors $\underline{x}^{(i)}$ 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 ~~defined~~ found by taking each node k and finding the minimum length l_{kl} ~~to each~~ of a path to each other node l. 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 hierarchially - choosing nearest node whenever there is a choice



This shows a typical ranking.

(3) Calculate d_i and D just as on page G53.

An alternate univariate method is the Wald-Wolfowitz run test. We start as in §53 (i) 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

this labels →	2	2	1	
set to which	x	x	x	# runs = 2
point	2	1	2	3
belongs	x	x	x	
	1	2	2	2
	x	x	x	

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 distributions. The mean number of runs is

$$\underbrace{1}_{\text{first point}} + \sum_{k=2}^{N} \underbrace{\frac{N_1}{N}}_{k-1\text{'th type 1}} \underbrace{\frac{N_2}{N-1}}_{k\text{'th type 2}} + \underbrace{\frac{N_2}{N}}_{k-1\text{'th type 2}} \underbrace{\frac{N_1}{N-1}}_{k\text{'th type 1}}.$$

$N = N_1 + N_2$.

by counting chance of a change in set # at k'th point of pooled data.

$$\langle runs \rangle = 1 + \frac{2N_1N_2}{N_1+N_2}$$

$= 2\,1/3$ in example on §5.

We can clearly calculate standard deviation ~~and di~~ $\left(\sigma^2 = \frac{2N_1N_2(2N_1N_2-N)}{N^2(N-1)} \right)$ 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 than link ~~points~~ nodes (points $\underline{x}^{(n)}$) 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 $\{\underline{x}^{(i)}\ i=1..N\}$, we wish to estimate the underlying distribution $p(\underline{x})$. Here is a possible technique

(i) Transform co-ordinates to the principal axes so that the covariance matrix of the $\underline{x}^{(i)}$ is diagonal. We invent a first guess

$$t(\underline{x}) = t_1(x_1)\, t_2(x_2) \ldots t_d(x_d) \quad (*)$$

where $\underline{x}$ is a d dimensional vector. The product form (*) is consistent with a diagonal covariance matrix but the latter only implies (*) 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~~ (*) may be good enough - use the methods described in B to find out.

If not, try

$$p(\underline{x}) = t(\underline{x})\, r(\underline{x})$$

and form for each point $\underline{x}^{(i)}$ in sample

$$R = \int_{\substack{k\text{'th nearest} \\ \text{neighbour} \\ \text{volume } V_k(\underline{x}^{(i)})}} \frac{1}{t(\underline{x})}\, p_{actual}(\underline{x})\, d\underline{x}$$

where p_{actual} is actual probability distribution. If $p_{actual}(\underline{x}) = p(\underline{x})$

$$R = \int_{V_k(\underline{x}^{(i)})} r(\underline{x})\, d\underline{x} = V_k\, r(\underline{x}^{(i)})$$

where we have assumed that $r(\underline{x})$ is slowly varying over volume.

We deduce that

$$r(\underline{x}^{(i)}) = 1/V_k(\underline{x}^{(i)}) \cdot \frac{1}{N} \sum 1/t(\underline{x}^{(j)})$$

k+1 points $\underline{x}^{(j)}$ in k'th nearest neighbour volume to $\underline{x}^{(i)}$

Now we have associated a "response" $r(\underline{x}^{(i)})$ to each point $\underline{x}^{(i)}$. We can use the techniques described in §13 to smooth (parameterize) $r(\underline{x}^{(i)})$* and so we have now obtained our final estimate $t(\underline{x})\, r(\underline{x})$ to the underlying distribution. We did have to assume that $r(\underline{x})$ 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 parameterization for r positive!

§17 Multivariate Decisions

Suppose we want to classify a vector random variable $\underline{x}$ into one of two (you replace two by any integer) possibilities. We discussed the univariate case in §8. This problem crops up often e.g. you wish to classify a particle as either π or electron (here $\underline{x}$ represents momentum measurement and pulse heights in various parts of a calorimeter). Alternately you wish to classify a candidate track as either good or bad. Here $\underline{x}$ is such things as χ^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 <u>training vectors</u> 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 ~~separate~~ classify an arbitrary $\underline{x}$ as type I or II. Note that if $\int p^{(1)}(\underline{x})\, p^{(2)}(\underline{x})\, d\underline{x}$ is nonzero then there is an intrinsic ambiguity in at least some regions of $\underline{x}$ space

<u>A k'th Nearest Neighbour</u>

Given $\underline{x}$, find the k nearest neighbours to $\underline{x}$ from the pooled data $\{x^{(1,i)}, x^{(2,j)}\}$ 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 i.e. if $N_1 > N_2$ only count type I events as of weight N_2/N_1.

(ii) One should (a) choose k and (b) transform co-ordinates to change metric for nearest neighbour distance calculation; by requiring these choices to minimize misclassification of vectors in training samples i.e. to minimize total # 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) initialization 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

J. H. Friedman IEEE T of C p 404 April 1977.

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. $(F_1(x^*) > F_2(x^*))$. 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

$$D_{12}(x^*) = |F_1(x^*) - F_2(x^*)| .$$

We now describe Friedman's multivariate generalization of this.

Look at each component x_j of $\underline{x}$ and find $D_j = \max \{ F_I(x_j) - F_{II}(x_j) \}$ Here the cumulative distributions are estimated as in §3 from training sample. Choose j with maximum D_j. From previous page, a decision based on this j will have the smallest misclassification. The cut $x \lessgtr x_j^*$ 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 (i) 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 training vectors one can use it to make a very fast decision on any given $\underline{x}$ as to its category. This is described for π/e case in a memo by J. Dorfan (Mark II Internal note, 1979).