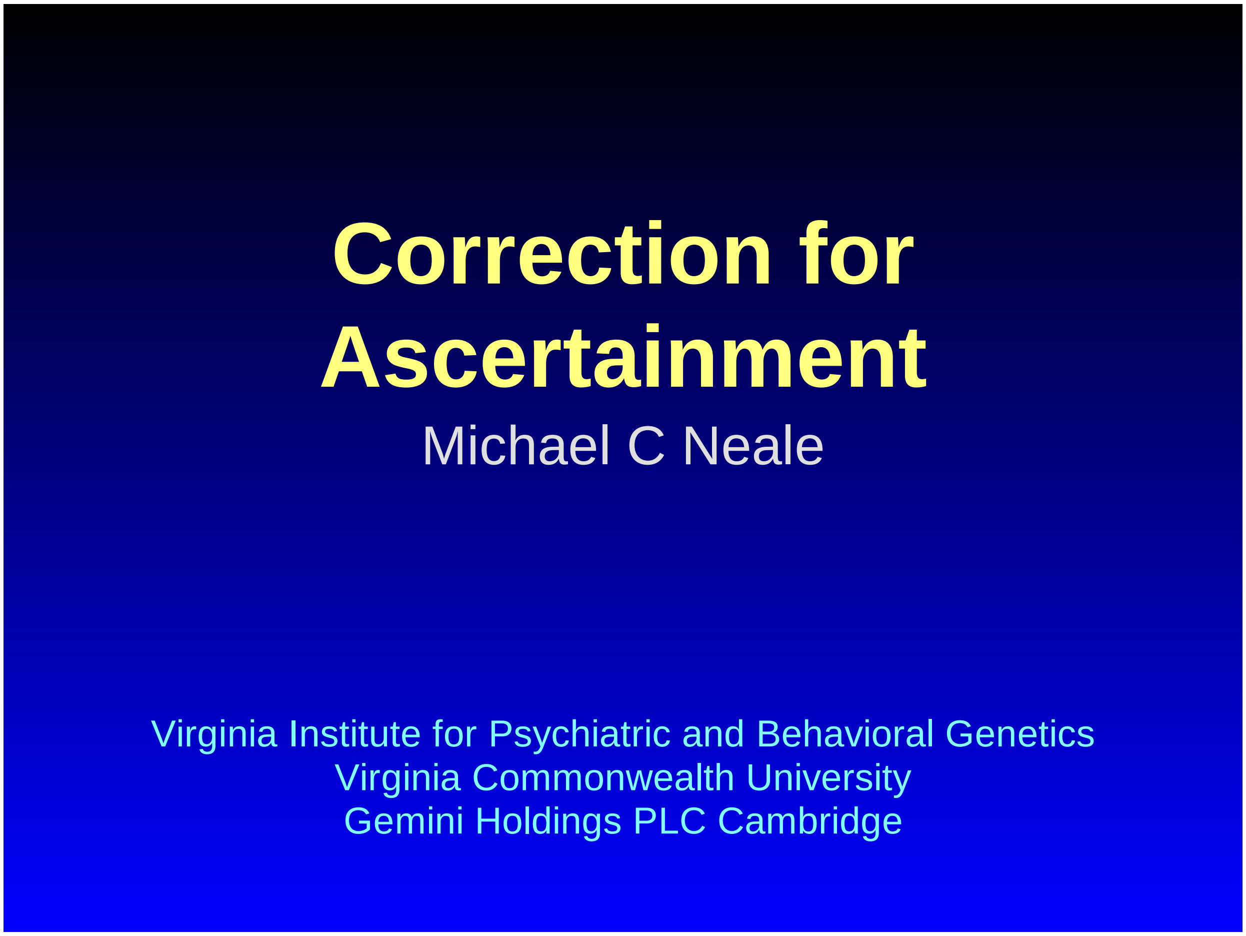
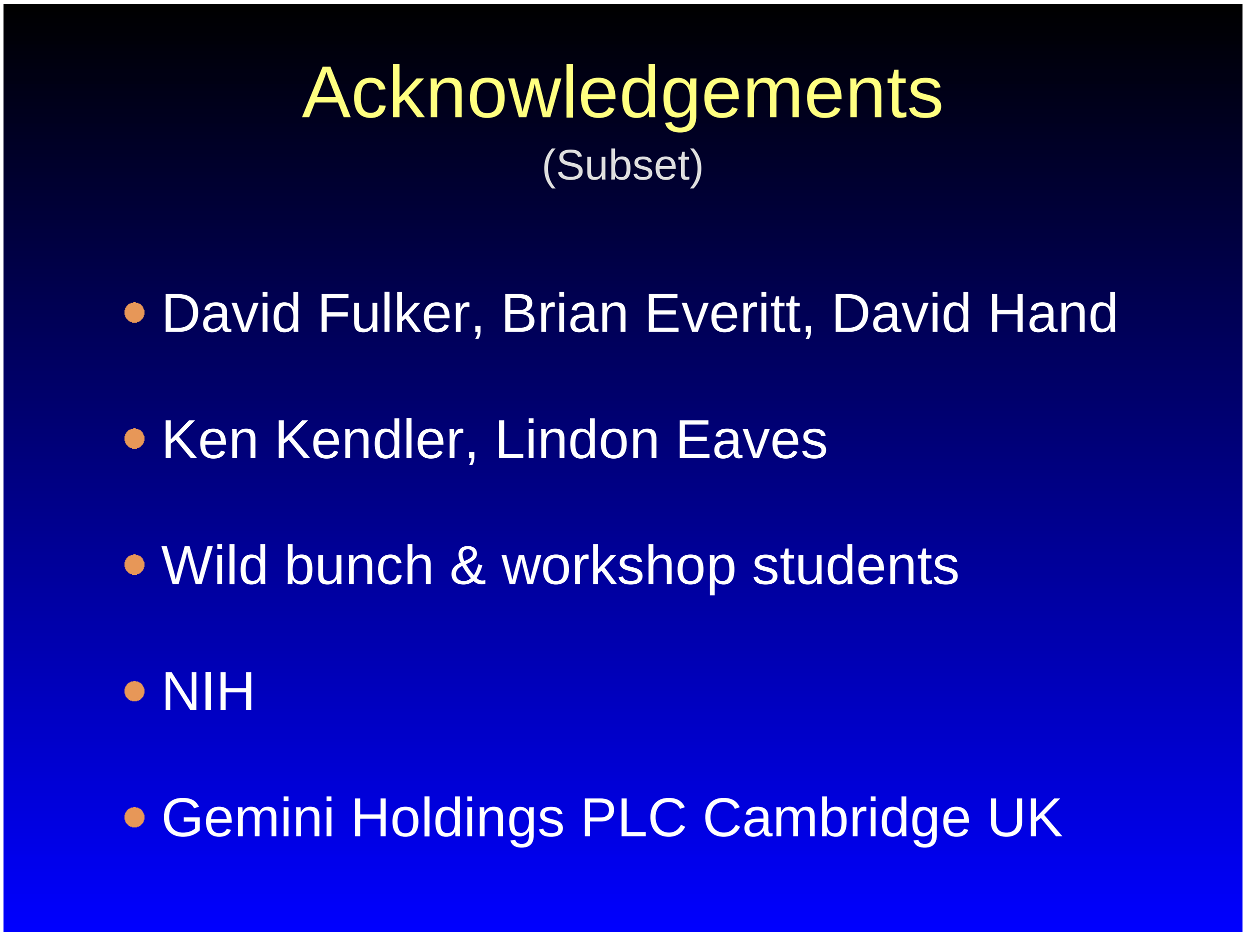


Correction for Ascertainment

Michael C Neale

Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University
Gemini Holdings PLC Cambridge



Acknowledgements

(Subset)

- David Fulker, Brian Everitt, David Hand
- Ken Kendler, Lindon Eaves
- Wild bunch & workshop students
- NIH
- Gemini Holdings PLC Cambridge UK

Rationale

Why use non-random ascertainment

- Statistical Power
- IBD 2 vs IBD 0 contrast
- Increase proportion of IBD 2's: ASP
- Increase proportion of IBD 1's: DSP
- Both: EDAC

Overview

- Rationale
- Normal Theory Maximum Likelihood
  - pros & cons
- Missing Data
- Correction for ascertainment

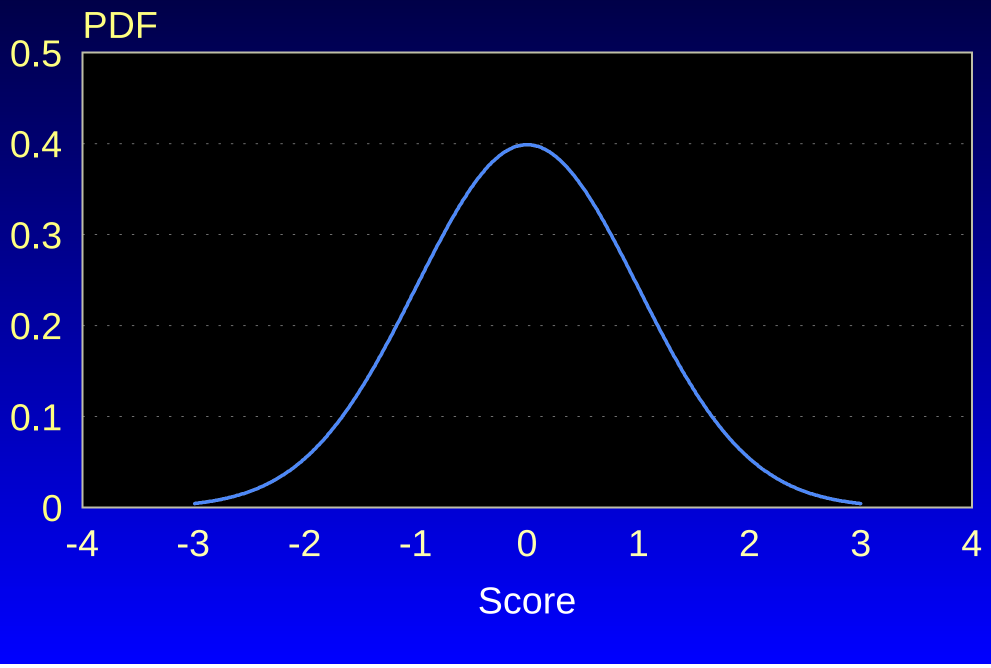
Maximum Likelihood Estimates

Have nice properties

- Asymptotically unbiased
- Minimum variance of all asymptotically unbiased estimators
- Invariant to transformations

Central Limit Theorem

Infinite factors of equal and small effect



Normal Theory Likelihood Function

For raw data in Mx

$$\ln L_i = f_i \sum_{j=1}^{m} \ln [w_j \;\; g(x_i,\mu_{ij},\Sigma_{ij})]$$

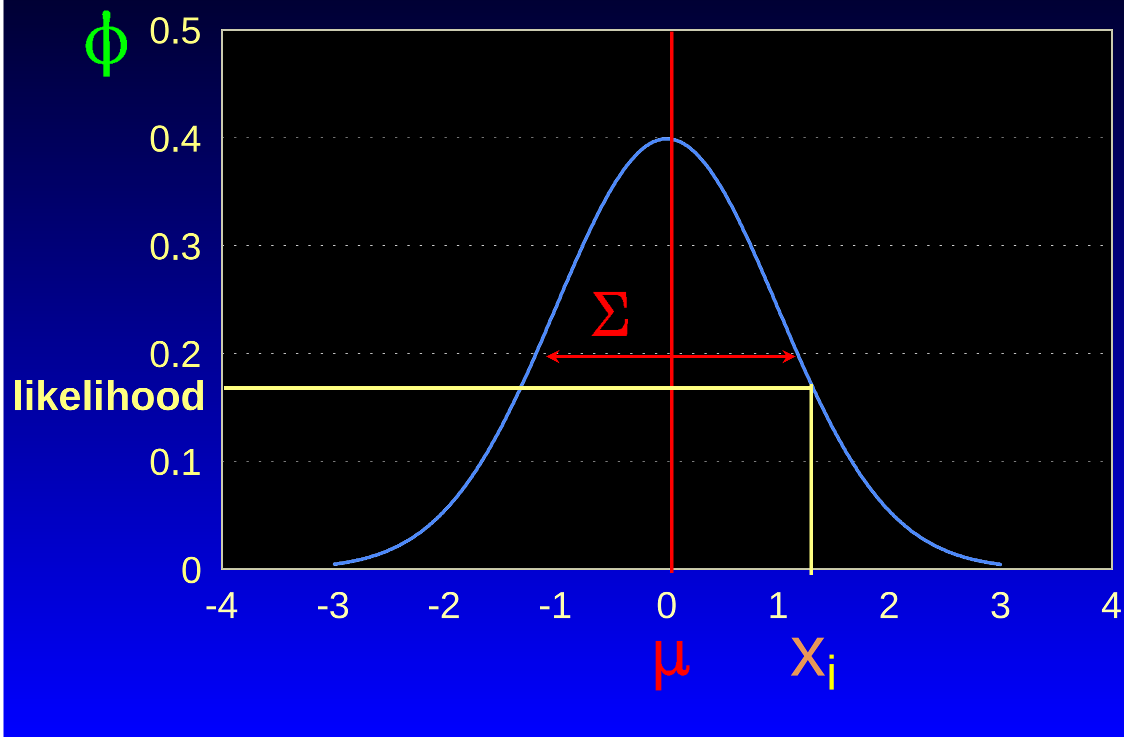
x_i - vector of observed scores
on n subjects

μ_{ij} - vector of predicted means

Σ_{ij} - matrix of predicted covariances
- functions of parameters

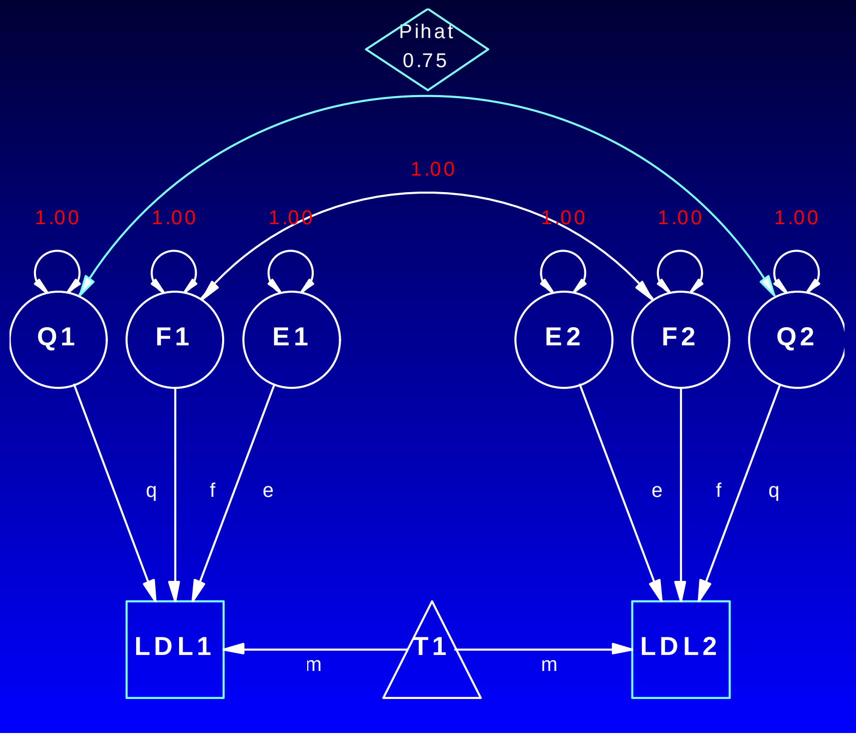
Normal distribution $\phi(\mu_{ij},\Sigma_{ij})$

Likelihood is height of the curve



Pihat Linkage Model for Siblings

Each sib pair i has different COVARIANCE



Weighted mixture of models

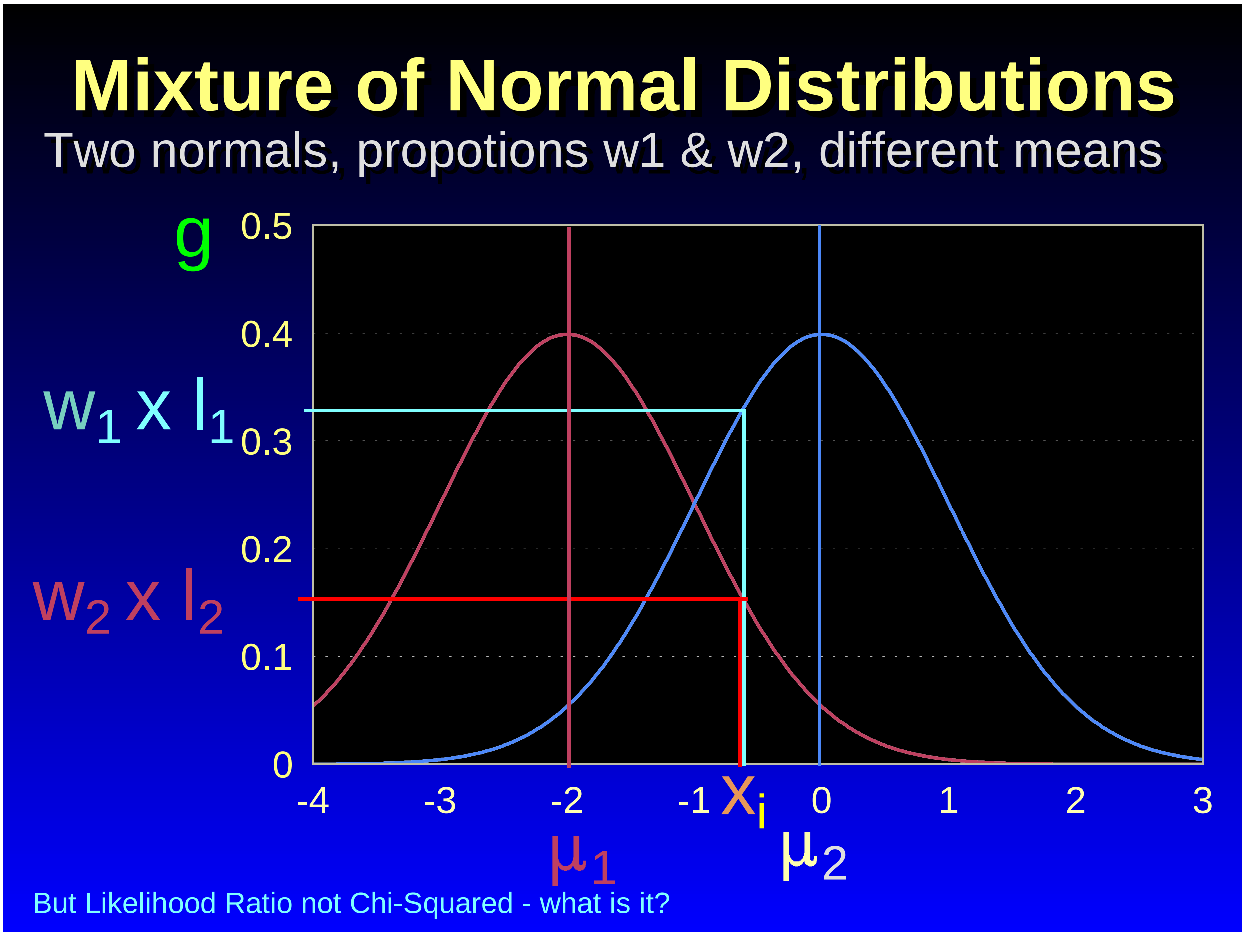
Finite mixture distribution

$$\ln L_i = f_i \sum_{j=1}^{m} \ln [w_{ij} \quad g(x_i, \mu_{ij}, \Sigma_{ij})]$$

$j = 1....m$ models

w_{ij} Weight for subject i model j

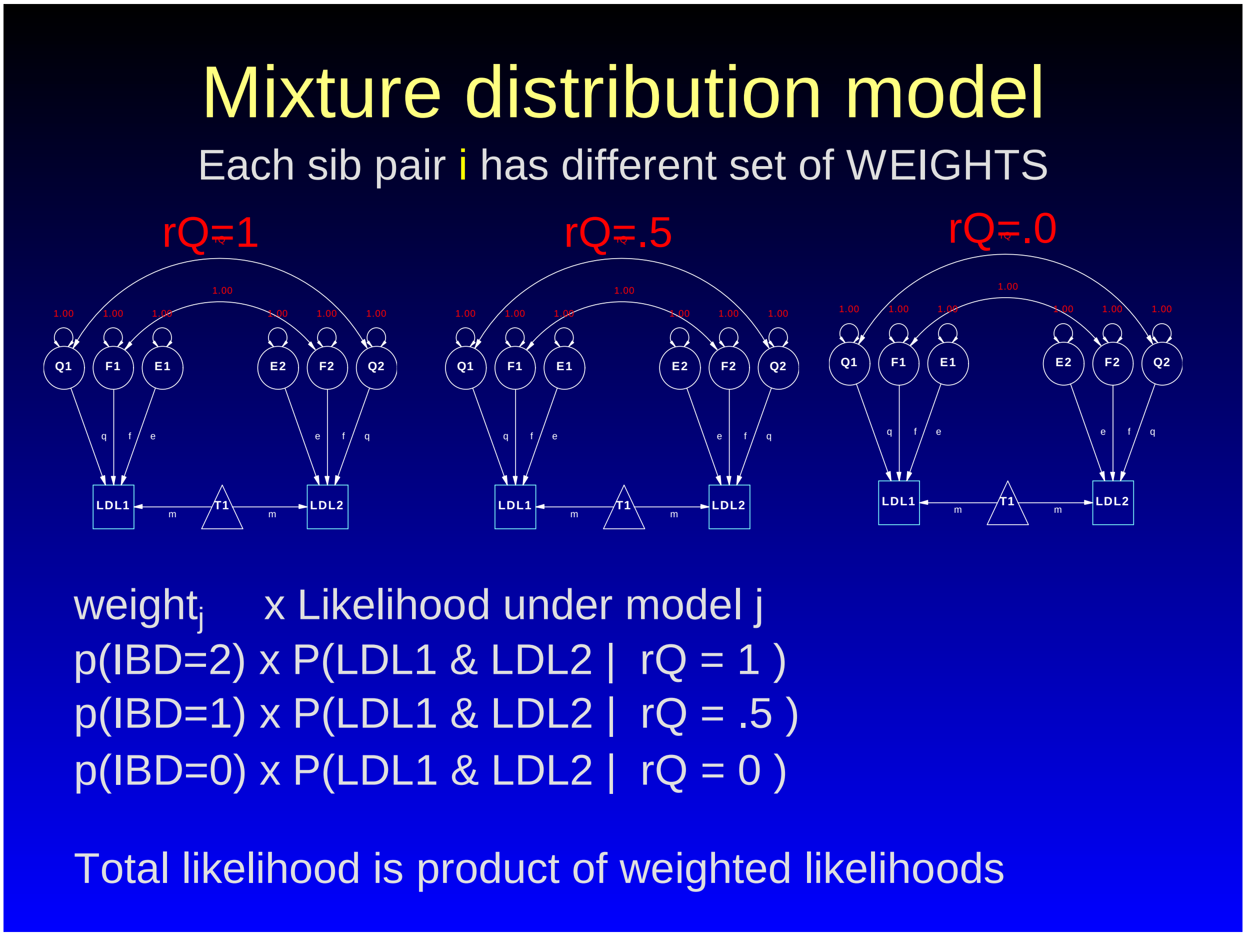
e.g., Segregation analysis



Mixture of Normal Distributions

Two normals, propotions w1 & w2, different means

But Likelihood Ratio not Chi-Squared - what is it?



Mixture distribution model

Each sib pair i has different set of WEIGHTS

weight$_j$ x Likelihood under model j

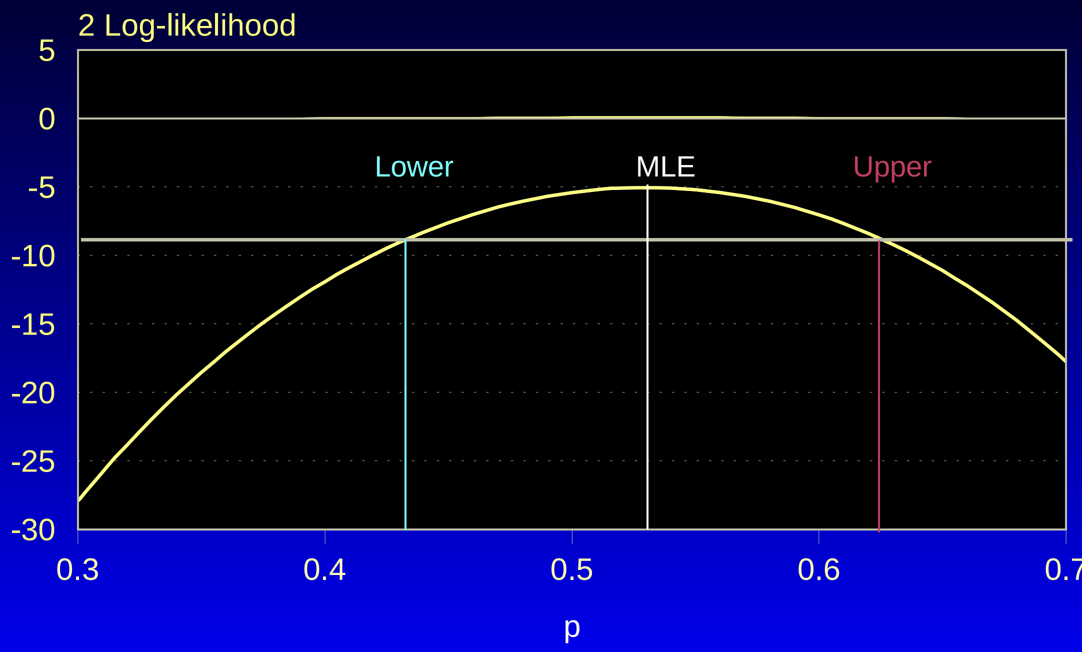
p(IBD=2) x P(LDL1 & LDL2 | rQ = 1)

p(IBD=1) x P(LDL1 & LDL2 | rQ = .5)

p(IBD=0) x P(LDL1 & LDL2 | rQ = 0)

Total likelihood is product of weighted likelihoods

Likelihood-based confidence interval



3.84 units of 2*ln L give 95% confidence interval of approximately (.44; .63)

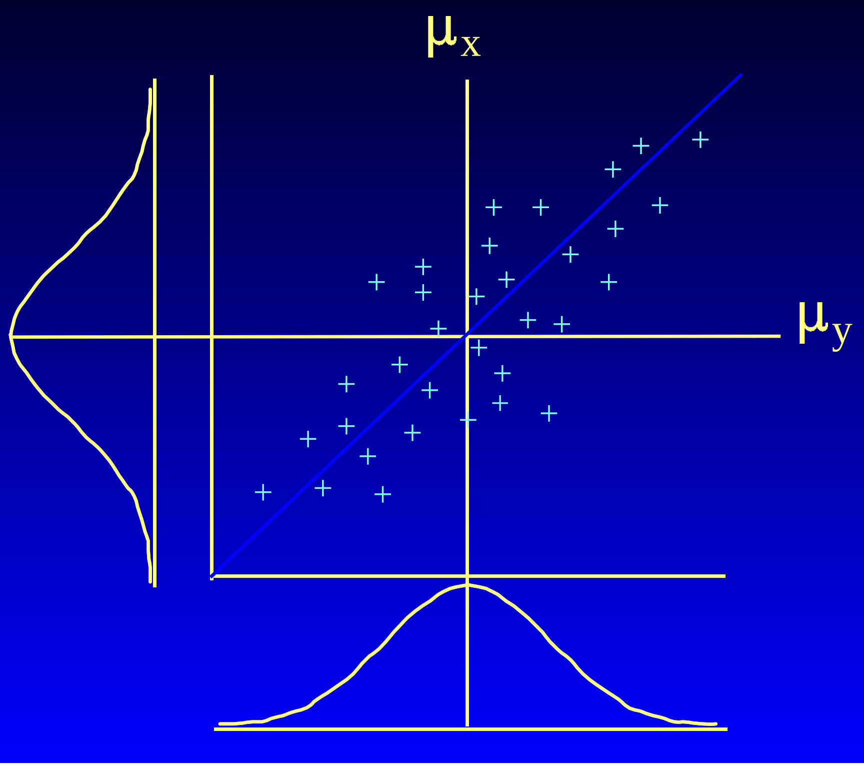
Computing Likelihood Based Confidence Intervals

- Fix parameter in question at successive values and maximize wrt rest (grid search)
- Plot graph and interpolate (spline search)
- Redefine fit function to be e.g.
  - (3.84 + Original fit)^2 +/- parameter value

Outlier detection

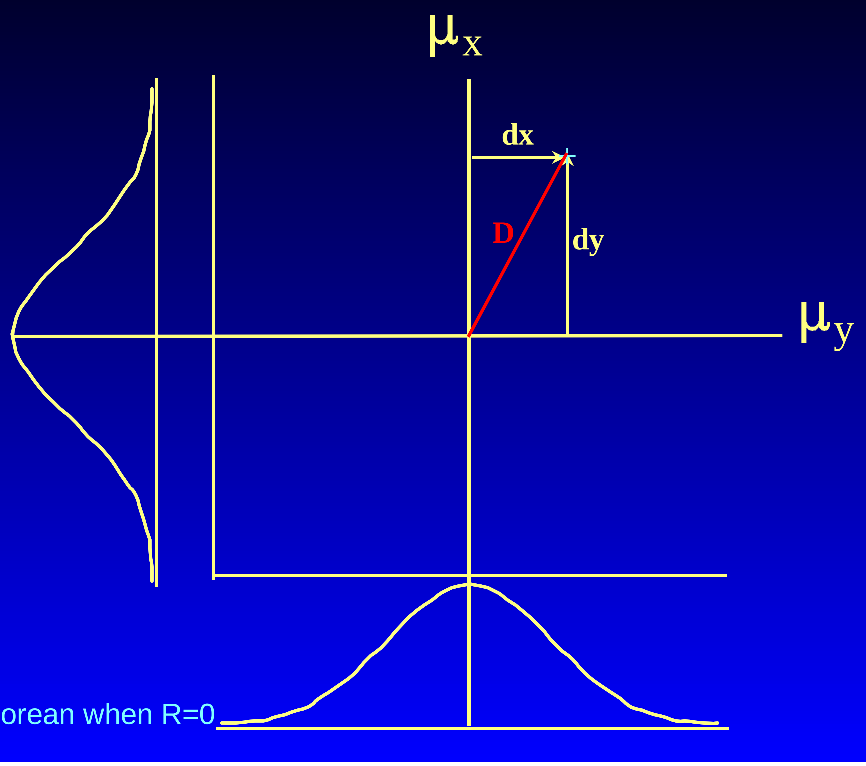
- Continuous data case
  - Mahalanobis distance
  - Z-score
- Can do something similar for Ordinal case
- Use option mx%p=filename to obtain individual fit statistics

Deviations in two dimensions



Deviations in two dimensions

Mahalanobis distance D



Pythagorean when R=0

Missing data

Little & Rubin 1987

- Missing completely at random
  - Causes of missingness independent
- Missing at random
  - Causes of missingness are either independent or measured
- Not missing at random
  - Due to residual variance in the missing variable itself

Computing likelihood

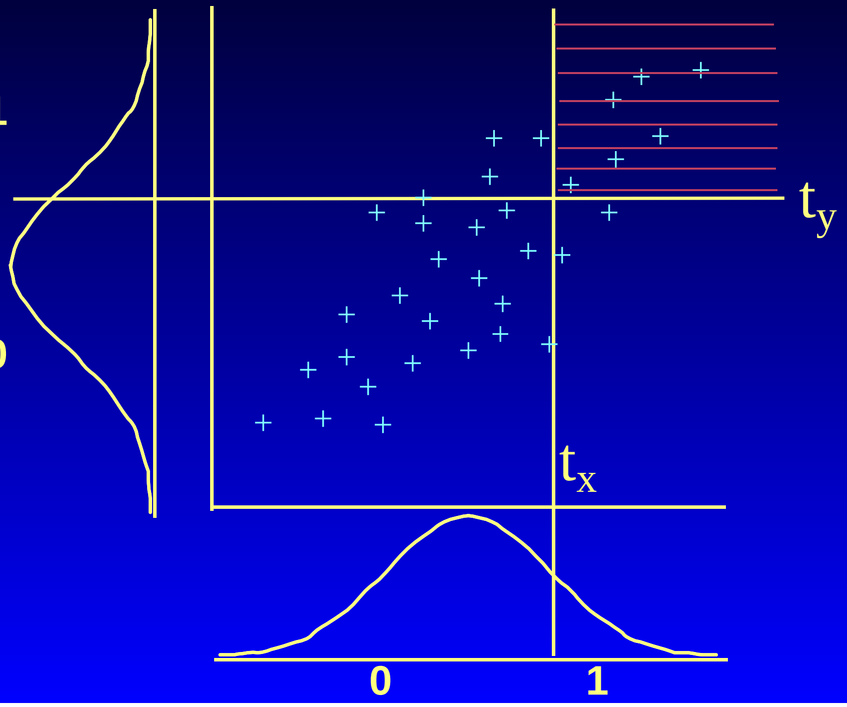
In presence of missing data

- Formally
  - Integrate over all missing value could be

$$\int_t^\infty \int_{-\infty}^\infty \phi(x,y)\,dx\,dy = \int_t^\infty \phi(y)\,dy$$

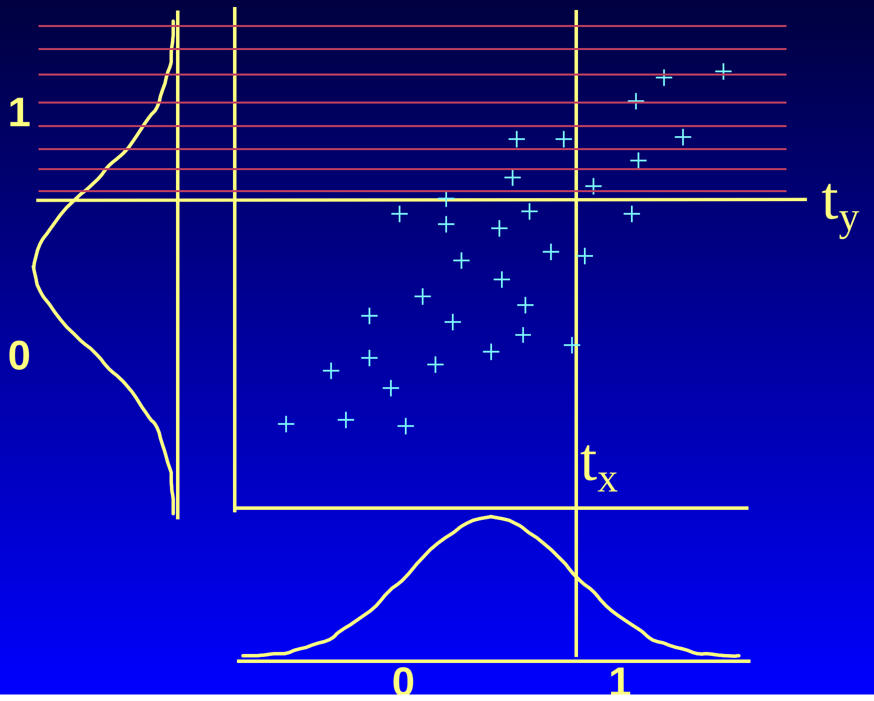
Data X = 1 Y = 1

$$\int_{tx}^\infty \int_{ty}^\infty \phi(x,y)\,dy\,dx$$



Data X = 1 Y = .

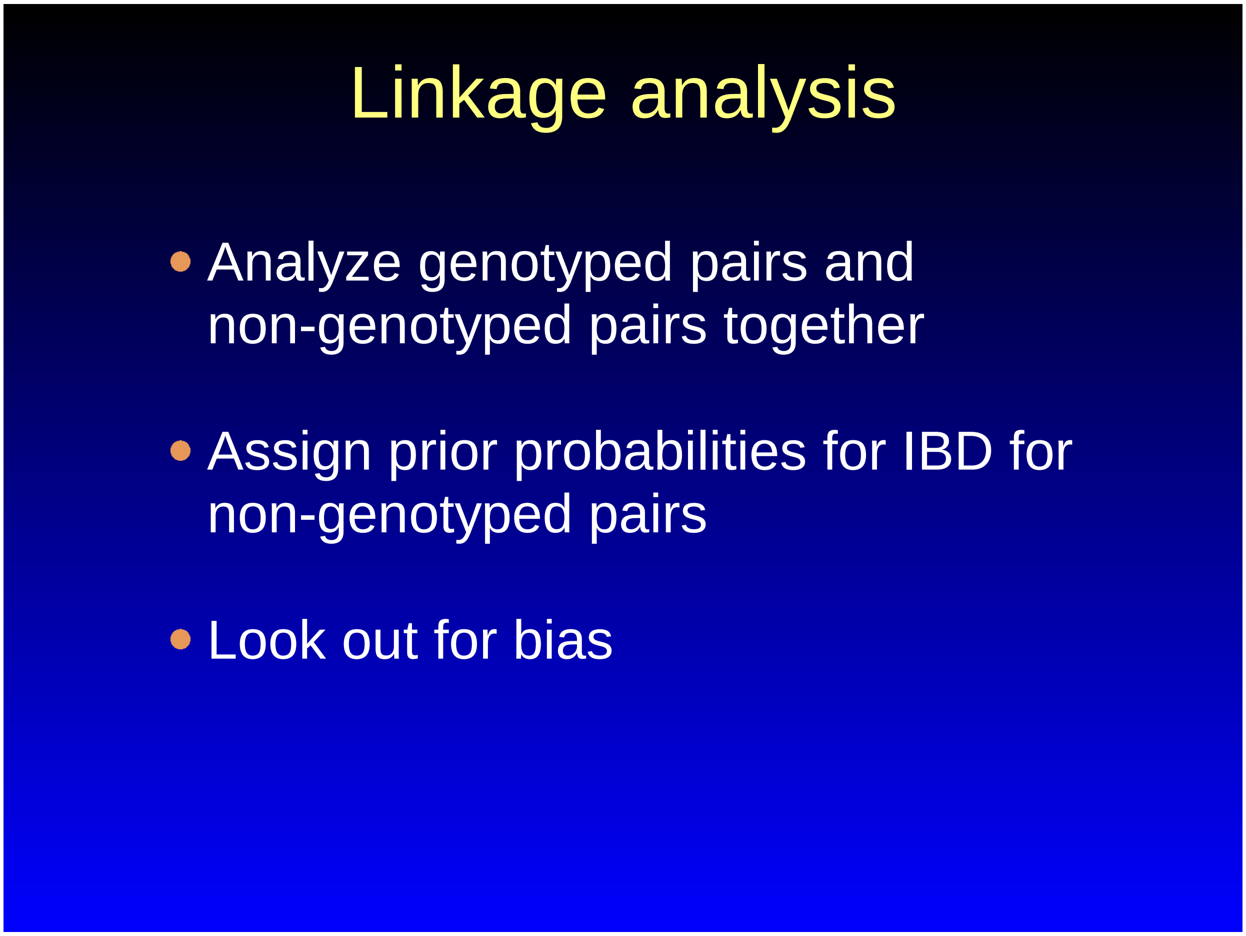
$$\int_{ty}^{\infty} \phi(y)\, dy = \int_{ty}^{\infty} \int_{-\infty}^{\infty} \phi(x,y)\, dx\, dy$$



In practice

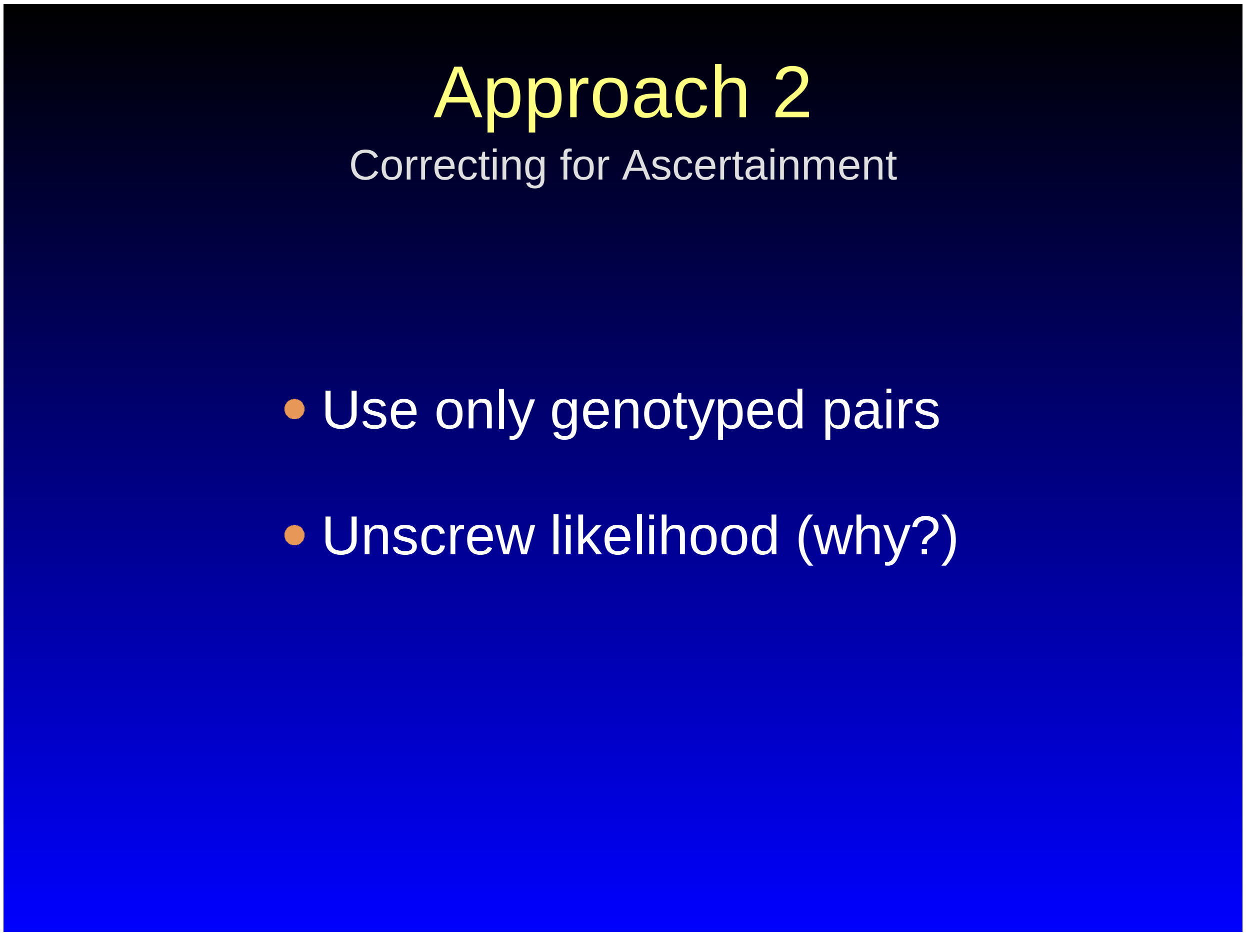
What Mx does

- Continuous case
  - Filter covariance and mean/threshold matrix and pretend
- Ordinal case
  - Filter threshold and covariance matrix and compute easier integral



Linkage analysis

- Analyze genotyped pairs and non-genotyped pairs together
- Assign prior probabilities for IBD for non-genotyped pairs
- Look out for bias



Approach 2

Correcting for Ascertainment

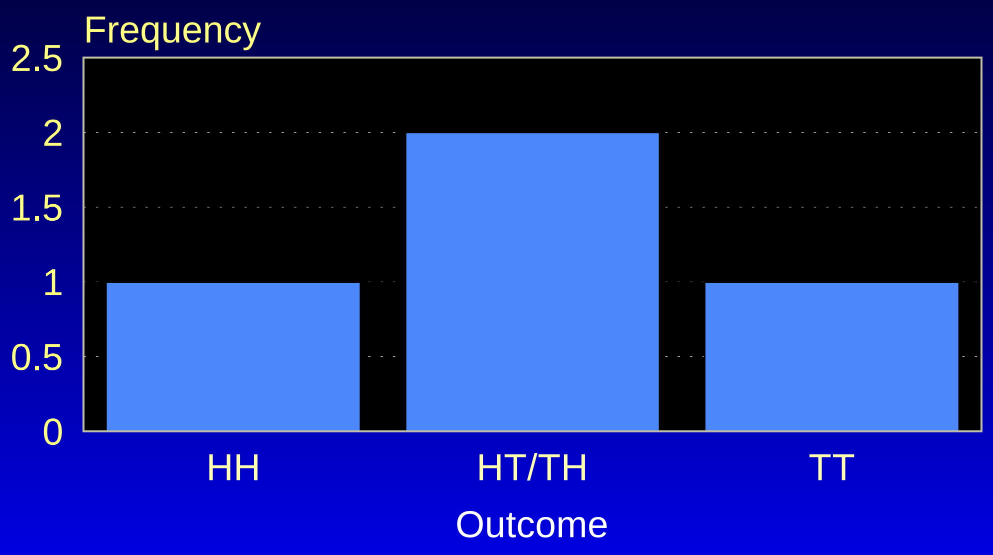
- Use only genotyped pairs
- Unscrew likelihood (why?)

Ascertainment Examples

- Studies of patients and controls
- Patients and relatives
- Linkage studies
  - Affected sib pairs, DSP etc
  - Multiple affected families

Example: Two Coin Toss

3 outcomes



Probability i = freq i / sum (freqs)

Non-random ascertainment

Example

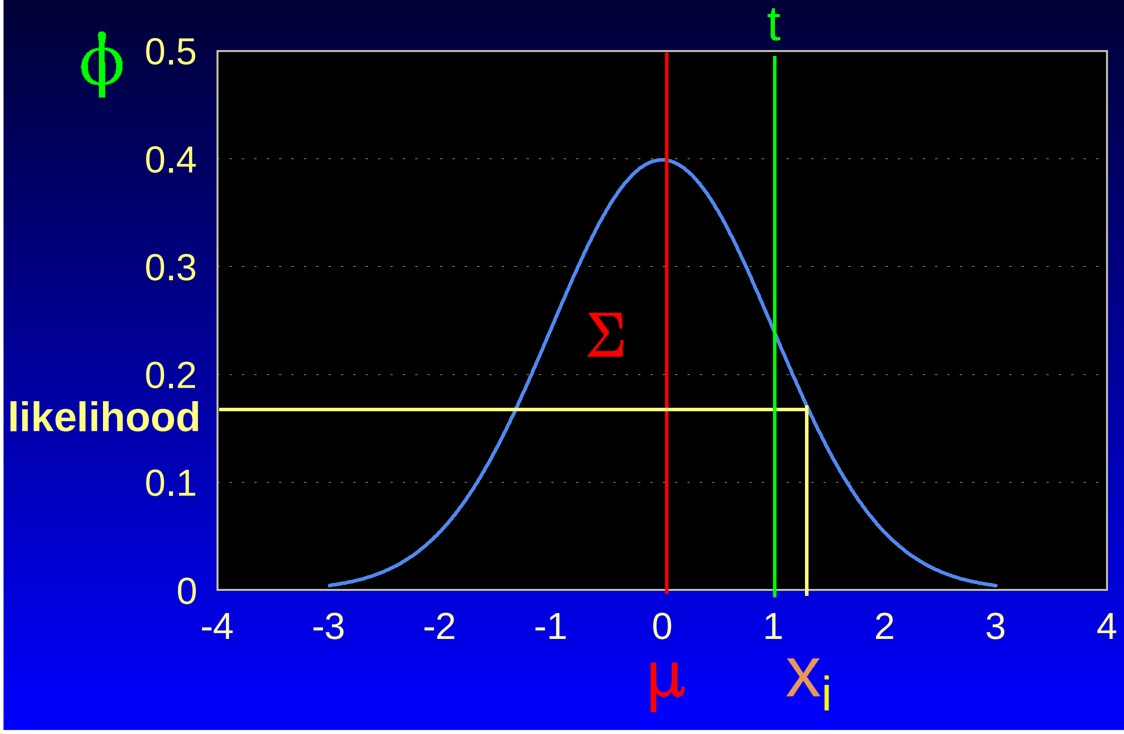
- Probability of observing TT globally
  - 1 outcome from 4 = 1/4

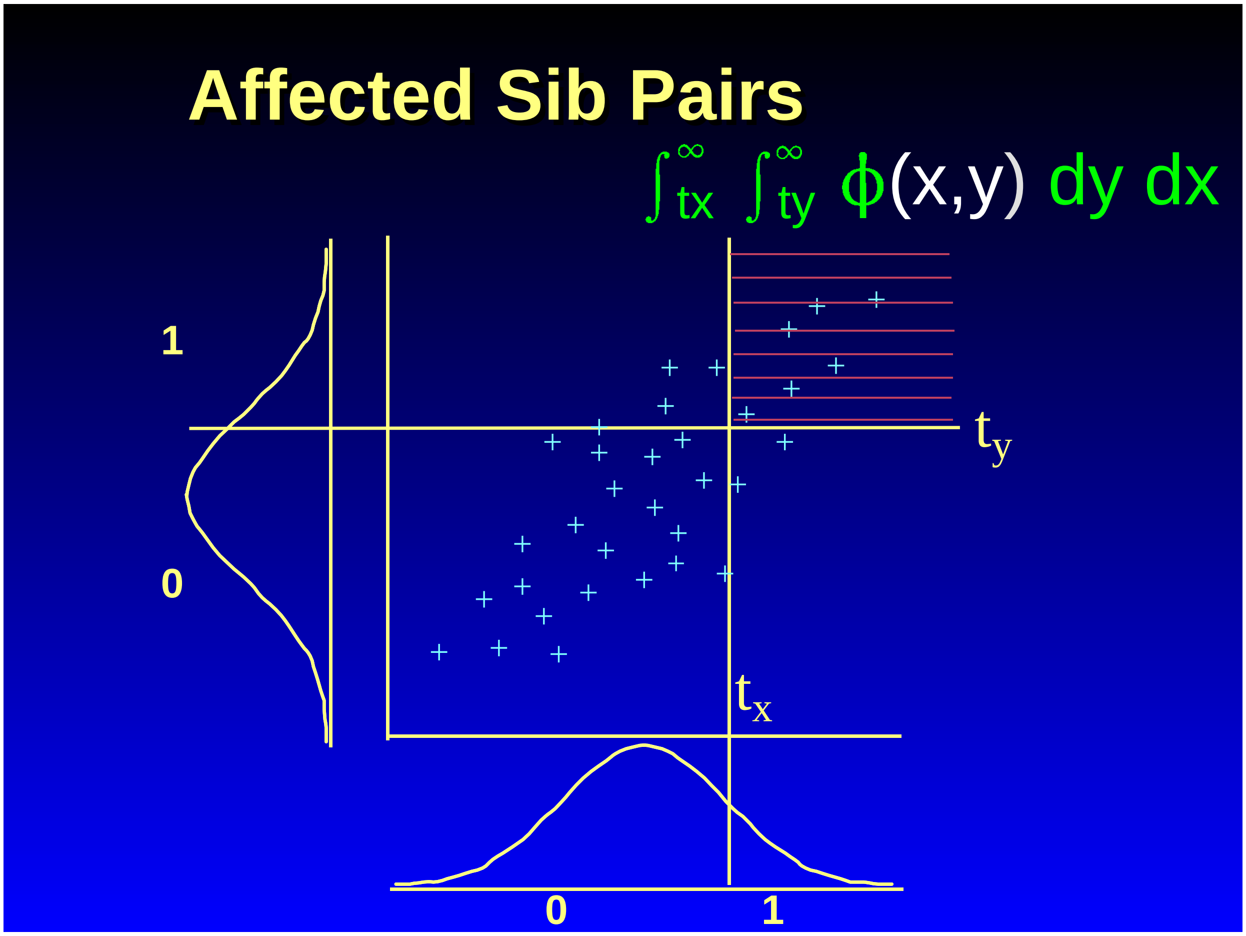
- Probability of observing TT if HH is not ascertained
  - 1 outcome from 3 = 1/3

  - or 1/4 divided by 'ascertainment correction' of 3/4 = 1/3

Correcting for ascertainment

Univariate case; only subjects > t ascertained





Affected Sib Pairs

$$\int_{tx}^{\infty} \int_{ty}^{\infty} \phi(x,y)\ dy\ dx$$

Correcting for ascertainment

Dividing by the realm of possibilities

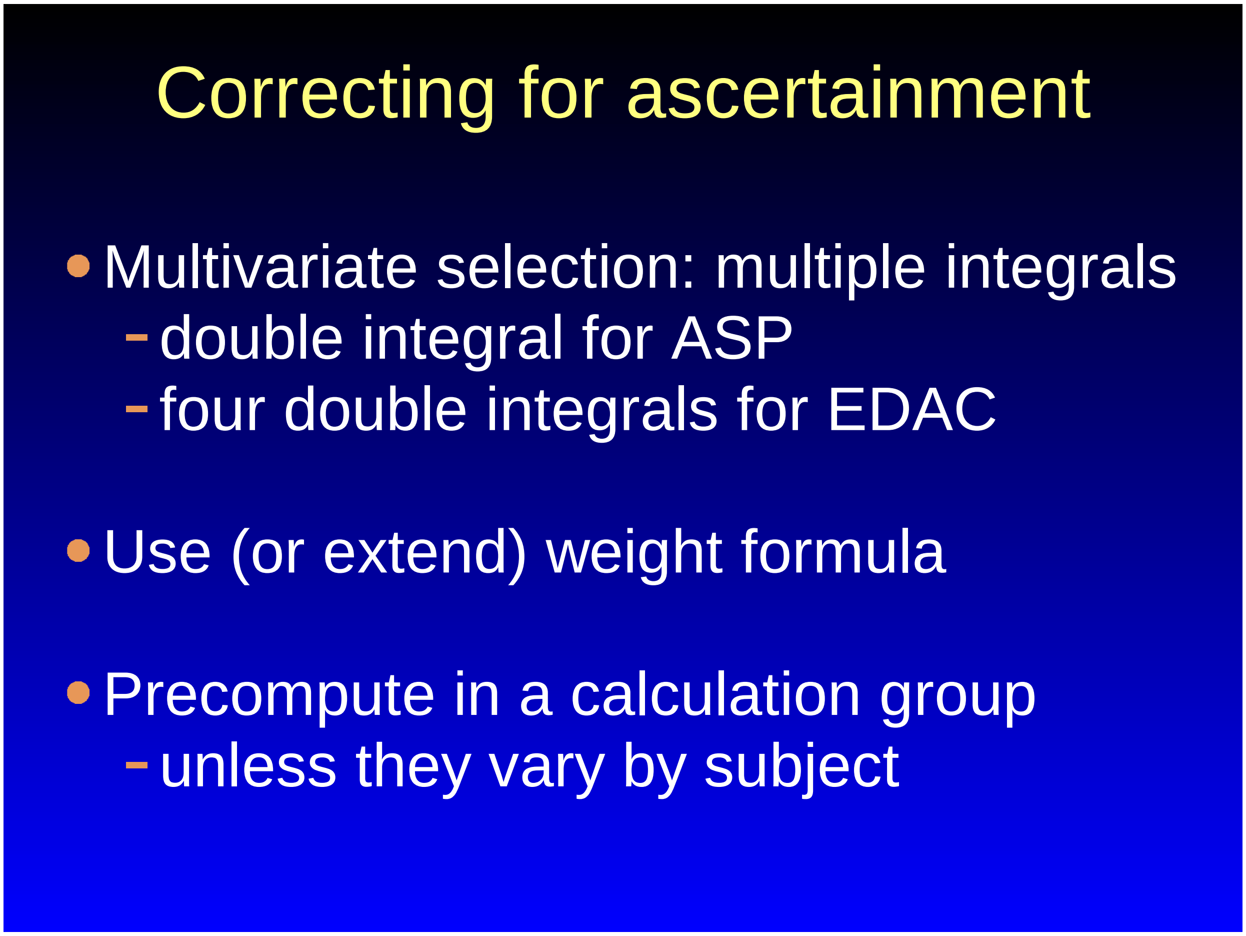
- Without ascertainment, we compute

  pdf, $\phi(\mu_{ij},\Sigma_{ij})$, at observed value x_i divided by:

  $$\int_{-\infty}^{\infty} \phi(\mu_{ij},\Sigma_{ij})\ dx \ = 1$$

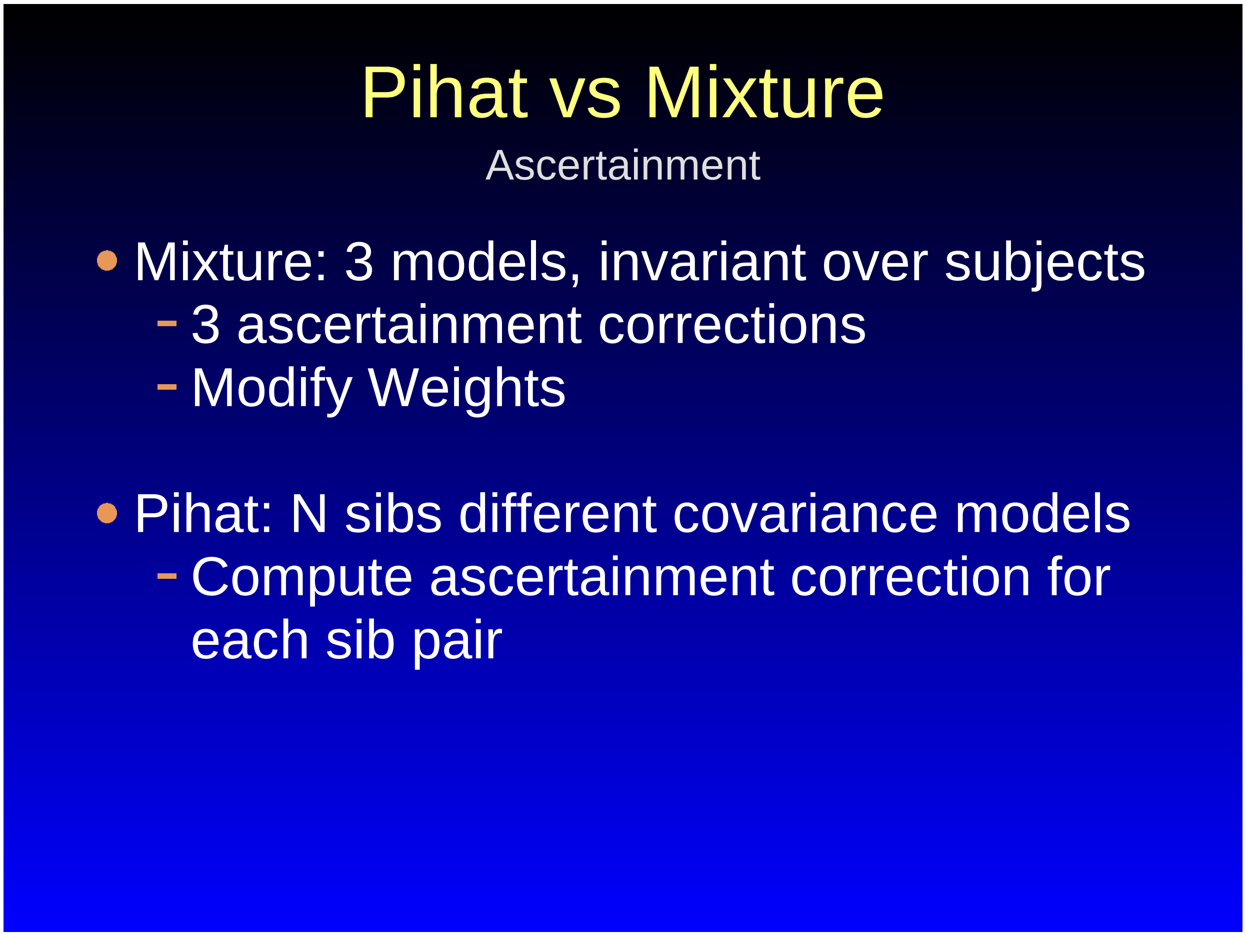
- With ascertainment, the correction is

  $$\int_{t}^{\infty} \phi(\mu_{ij},\Sigma_{ij})\ dx$$



Correcting for ascertainment

- Multivariate selection: multiple integrals
  - double integral for ASP
  - four double integrals for EDAC
- Use (or extend) weight formula
- Precompute in a calculation group
  - unless they vary by subject



Pihat vs Mixture

Ascertainment

- Mixture: 3 models, invariant over subjects
  - 3 ascertainment corrections
  - Modify Weights
- Pihat: N sibs different covariance models
  - Compute ascertainment correction for each sib pair

General Likelihood Function

What about the means μ_{ij}?

$$L_i = f_i \prod_{j=1}^{m} w_{ij} \quad g(x_i, \mu_{ij}, \Sigma_{ij})$$

Have varied Σ_{ij} (pihat) or w_{ij} (full IBD)

Association analysis varies μ_{ij}
causes trouble for asc correction

Correction for ascertainment

Joint linkage and association analysis

- Better watch out
- Correction w_j depends on
  - predicted means μ_{ij} (9 types)
  - predicted covariances, Σ_{ij} (3 types)
  - could still pre-compute 27 integrals & pick
- Careful if you are modeling covariates like age via means

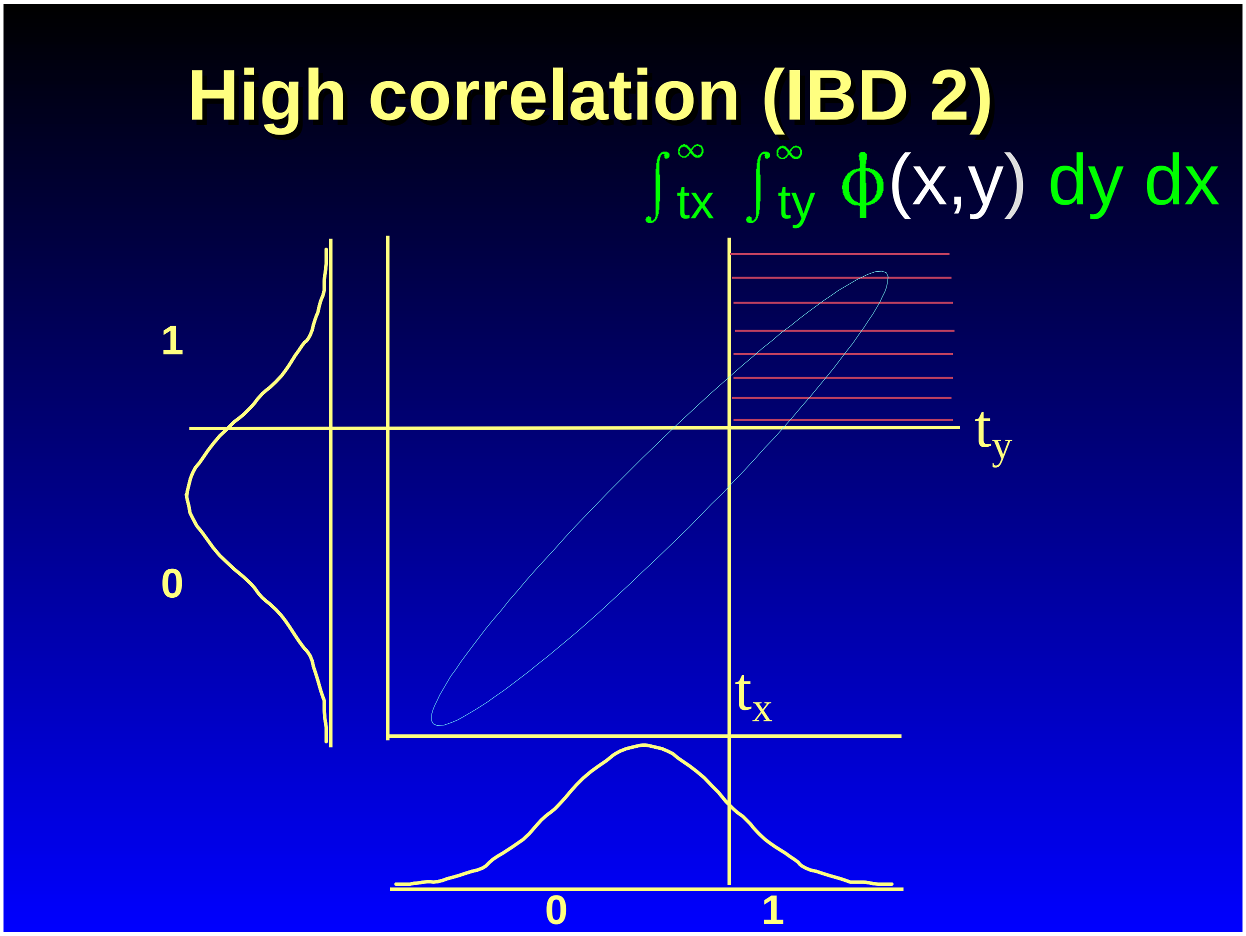
Two sources of information

In selected samples

- Difference in covariance as a function of IBD status
- Deviation of average pihat from .5
- Use them both?
  - Read in pihat in a separate group
  - Estimate mean & variance
  - Set mean to .5

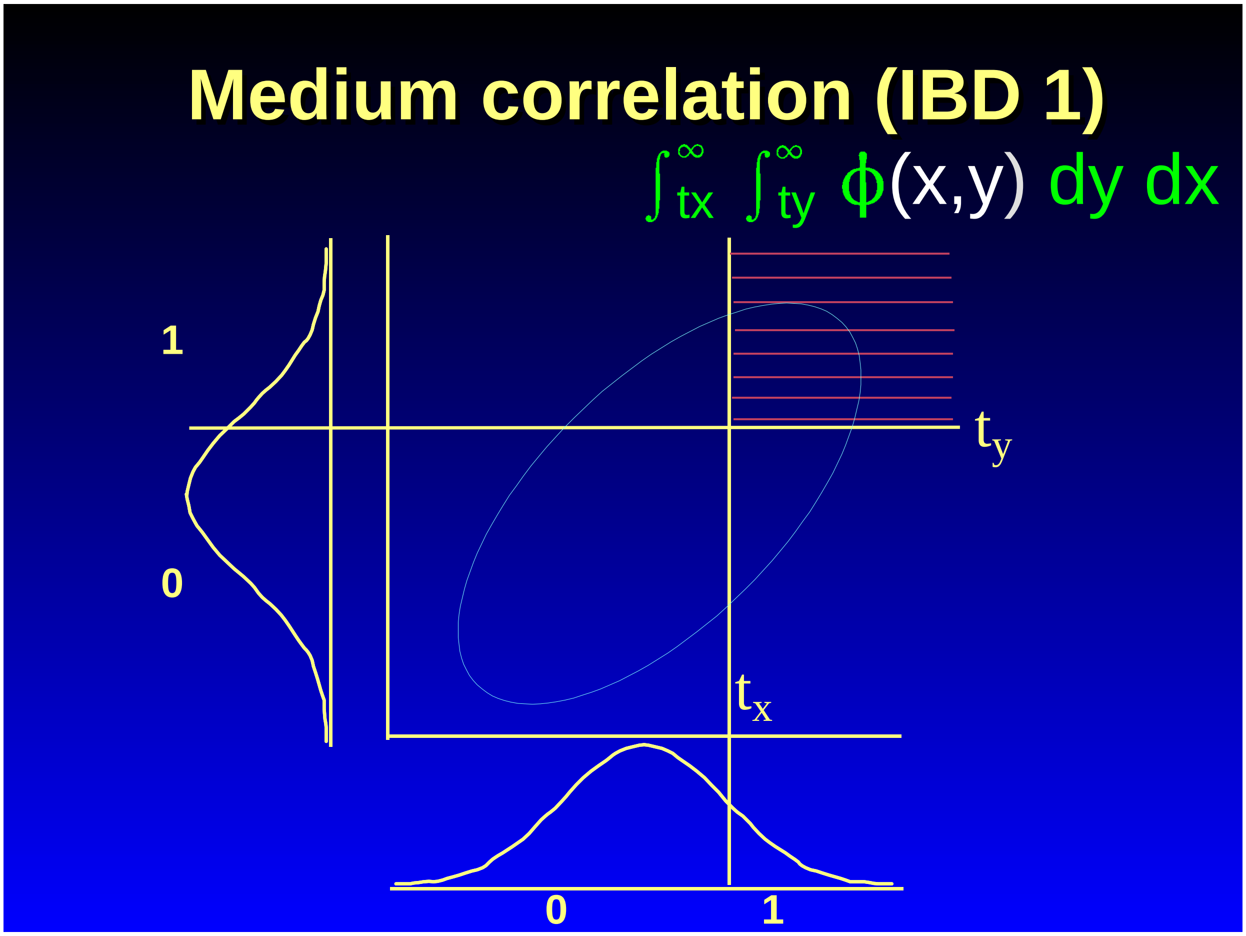
Expected Pihat Approach

- For a given q^2 can we predict what pihat should be under selection?
- Three distributions, initially .25 .5 .25
- Compute three integrals
  - recompute proportions



High correlation (IBD 2)

$$\int_{tx}^{\infty} \int_{ty}^{\infty} \phi(x,y)\ dy\ dx$$

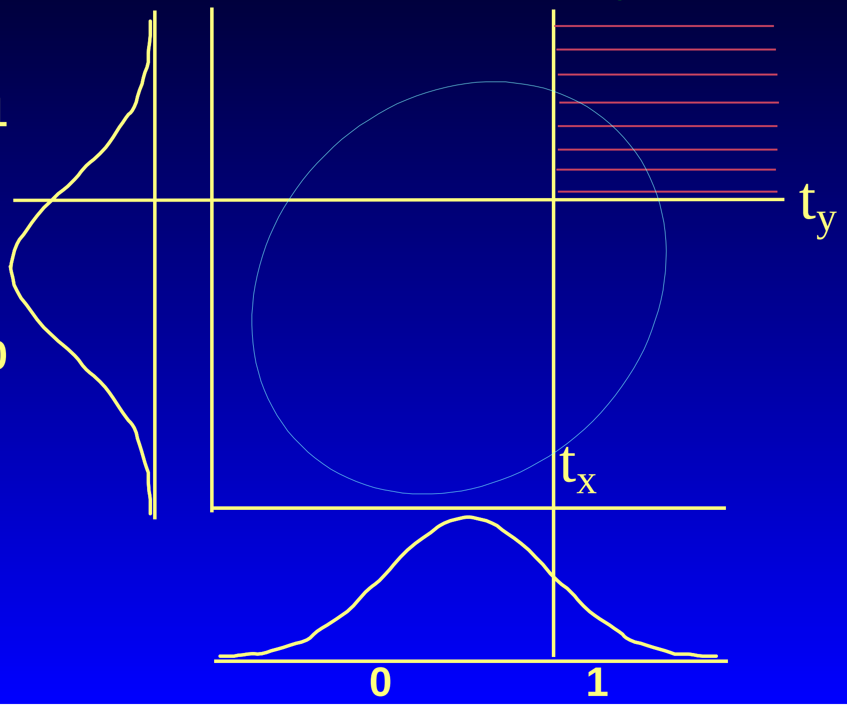


Medium correlation (IBD 1)

$$\int_{tx}^{\infty} \int_{ty}^{\infty} \phi(x,y)\ dy\ dx$$

Low correlation (IBD 0)

$$\int_{tx}^{\infty}\int_{ty}^{\infty}\phi(x,y)\,dy\,dx$$



Conclusion

- Can handle non-random ascertainment in two ways
- Include screened but not genotyped pairs in analysis
- Use only genotyped pairs
- Make use of 'marginal' average pihat info