VCU

Virginia Commonwealth University

Introduction to Linkage and Association for Quantitative Traits

Michael C Neale
Boulder Colorado Workshop March 2 2009

Overview

- A brief history of SEM
- Regression
- Maximum likelihood estimation
- Models
  - Twin data
  - Sib pair linkage analysis
  - Association analysis

Origins of SEM

- Regression analysis
  - 'Reversion' Galton 1877: Biological phenomenon
  - Yule 1897 Pearson 1903: General Statistical Context
  - Initially Gaussian X and Y; Fisher 1922 Y|X

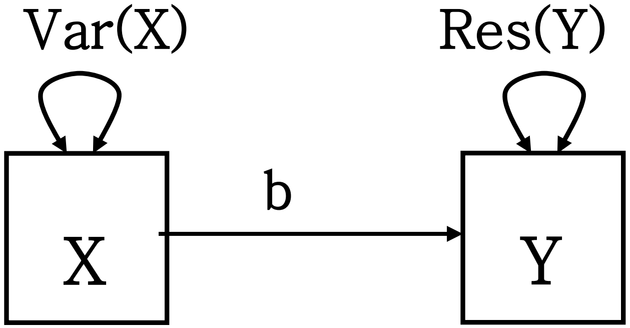
- Path Analysis
  - Sewall Wright 1918; 1921
  - Path Diagrams of regression and covariance relationships

Structural Equation Modeling Basics

- Two kinds of relationships
  - Linear *regression* X -> Y single-headed
  - Unspecified *covariance* X<->Y double-headed

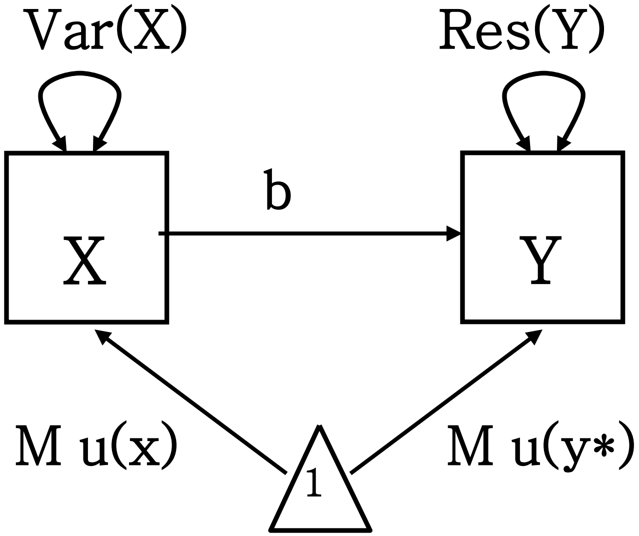
- Four kinds of variable
  - Squares: observed variables
  - Circles: latent, not observed variables
  - Triangles: constant (zero variance) for specifying means
  - Diamonds: observed variables used as moderators (on paths)

Linear Regression Covariance SEM



Models *covariances* only
Of historical interest

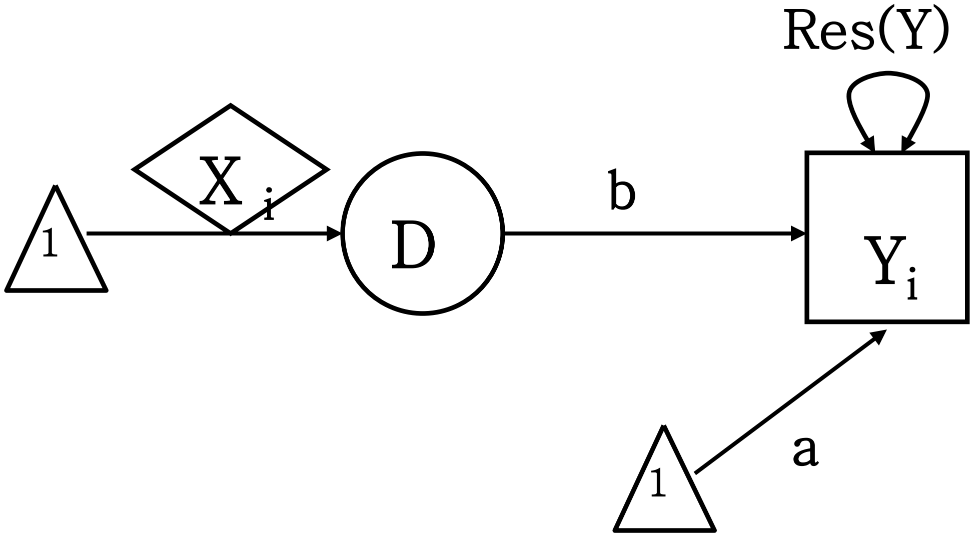
Linear Regression SEM with means



Models Means and Covariances

Linear Regression SEM: Individual-level

$Y_i = a + bX_i$



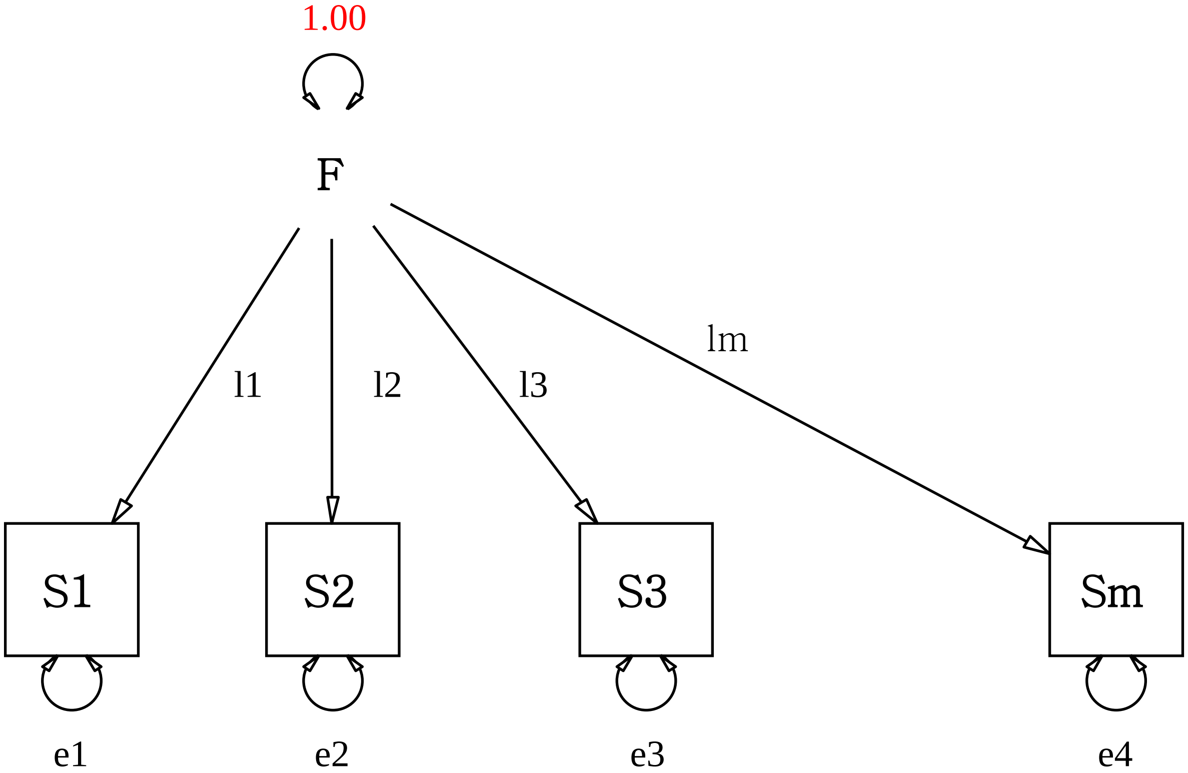
Models Mean and Covariance of Y *only*

Must have raw (individual level) data

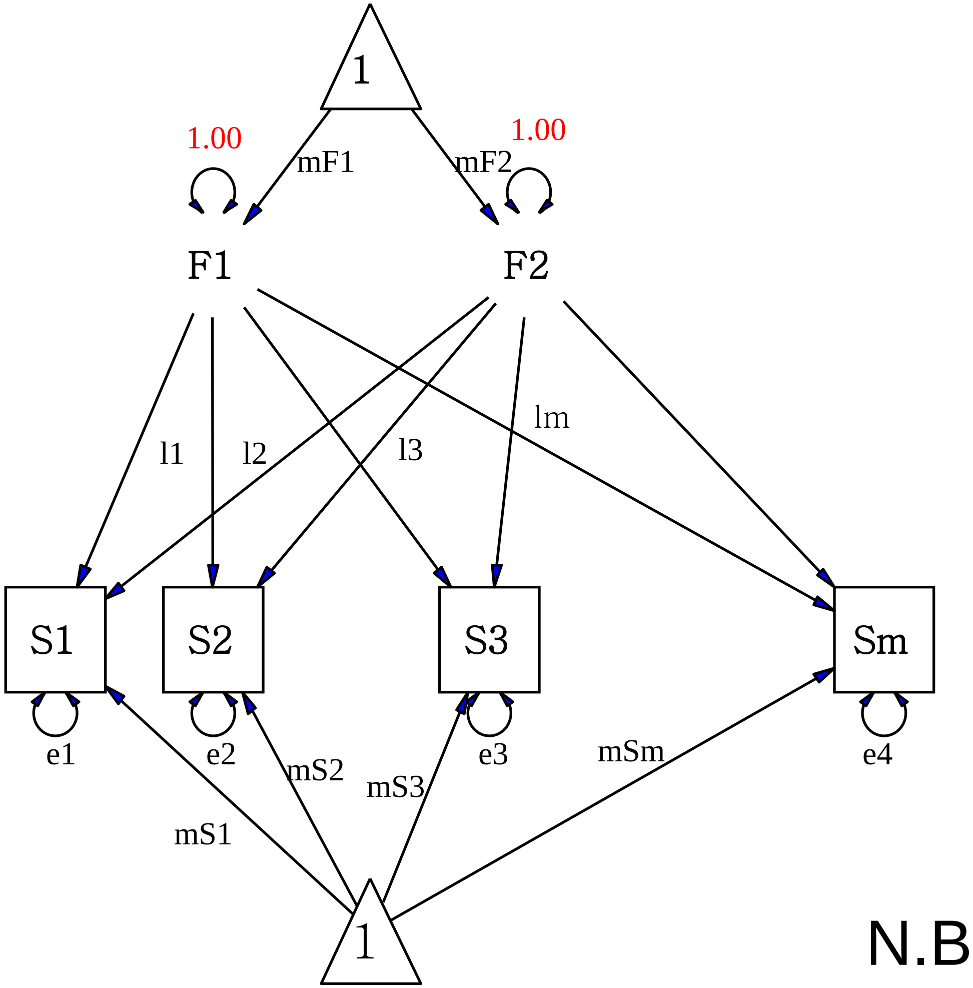
X_i is a *definition* variable

Mean of Y different for every observation

Single Factor Covariance Model



Two Factor Model with Covs & Means



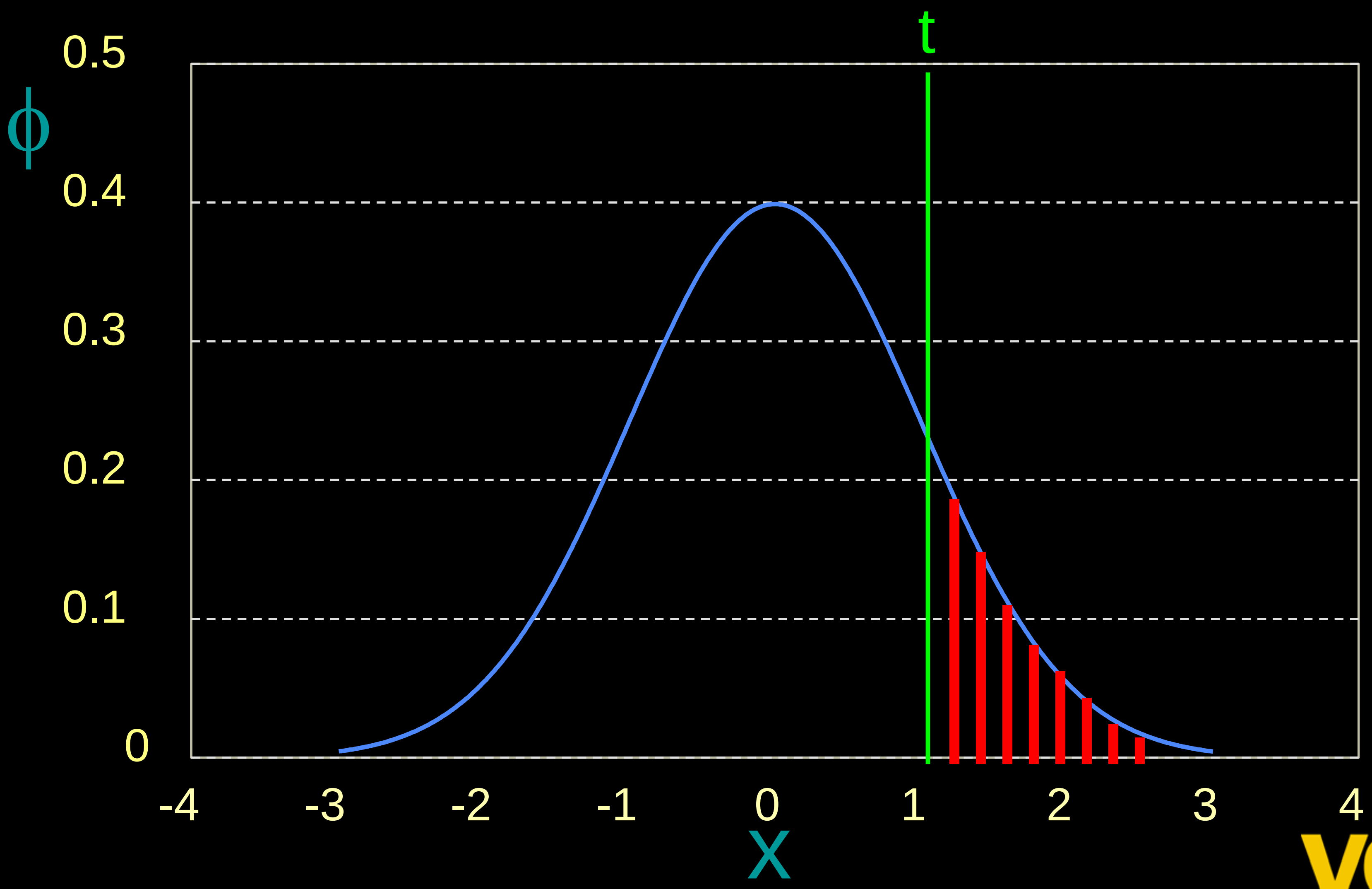
N.B. Not identified

Factor model essentials

- In SEM the factors are typically assumed to be normally distributed
- May have more than one latent factor
- The error variance is typically assumed to be normal as well
- May be applied to binary or ordinal data
  - Threshold model

Multifactorial Threshold Model

Normal distribution of liability. ‘Affected’ when liability x > t



Measuring Variation

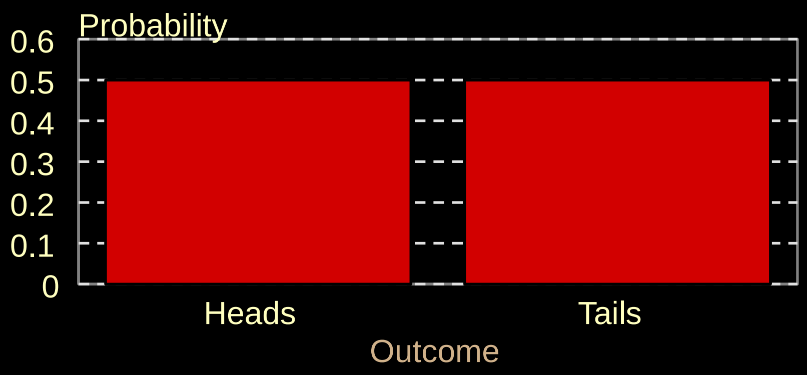
- Distribution
  - Population
  - Sample
  - Observed measures

- Probability density function ‘pdf’
  - Smoothed out histogram
  - f(x) >= 0 for all x

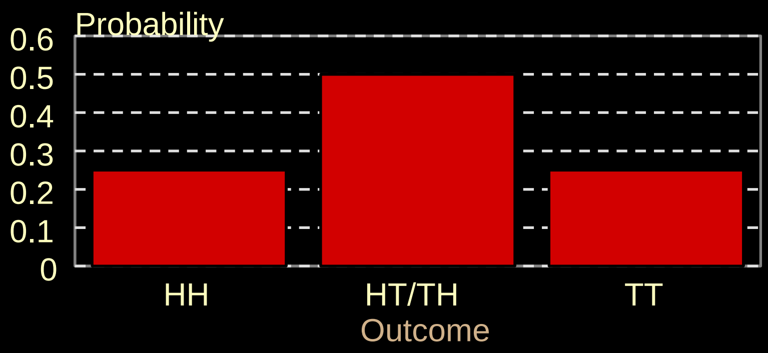
$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Flipping Coins

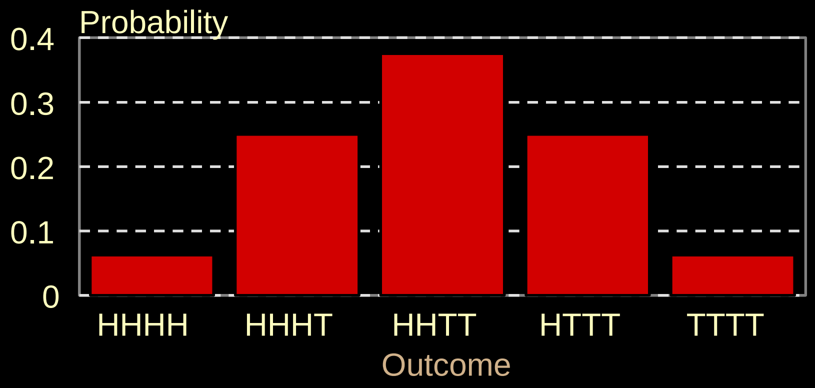
1 coin: 2 outcomes



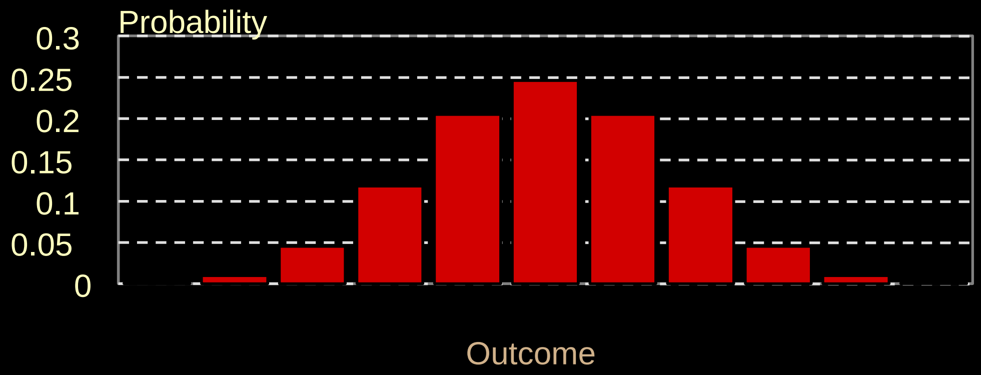
2 coins: 3 outcomes



4 coins: 5 outcomes

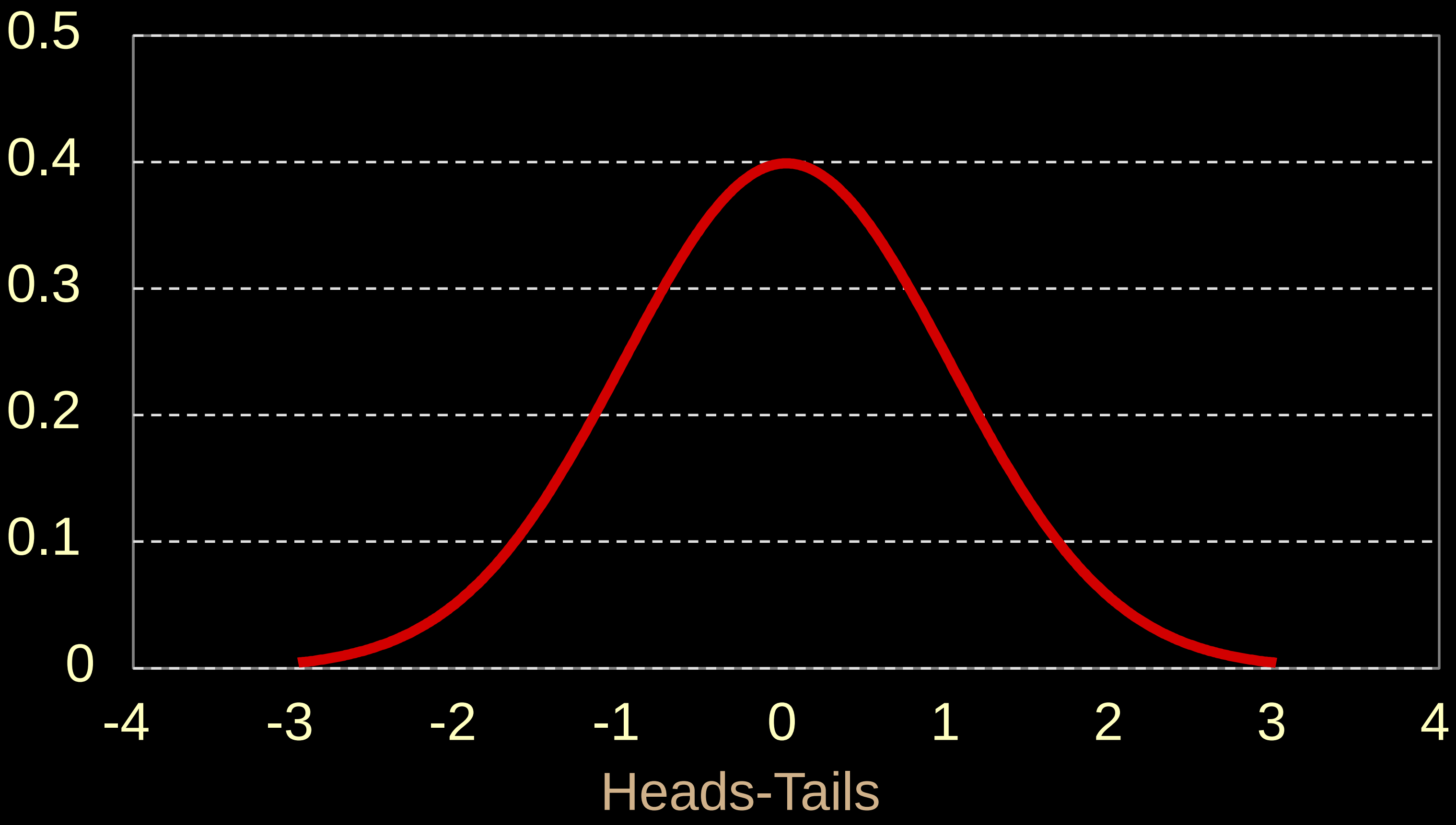


8 coins: 9 outcomes



Bank of China Coin Toss

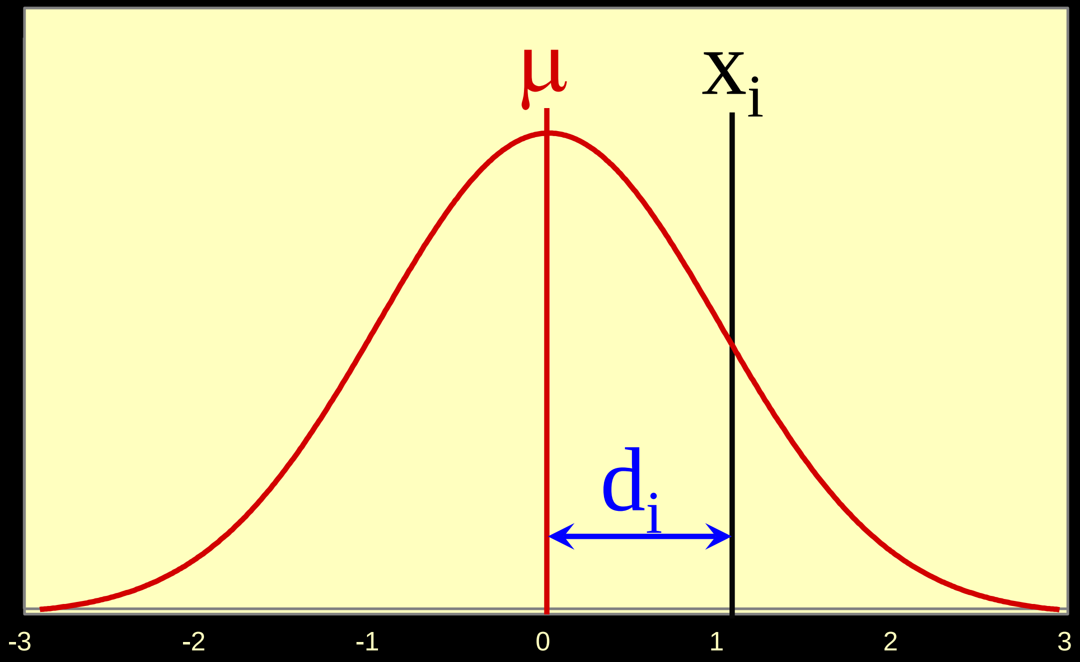
Infinite outcomes



De Moivre 1733 Gauss 1827

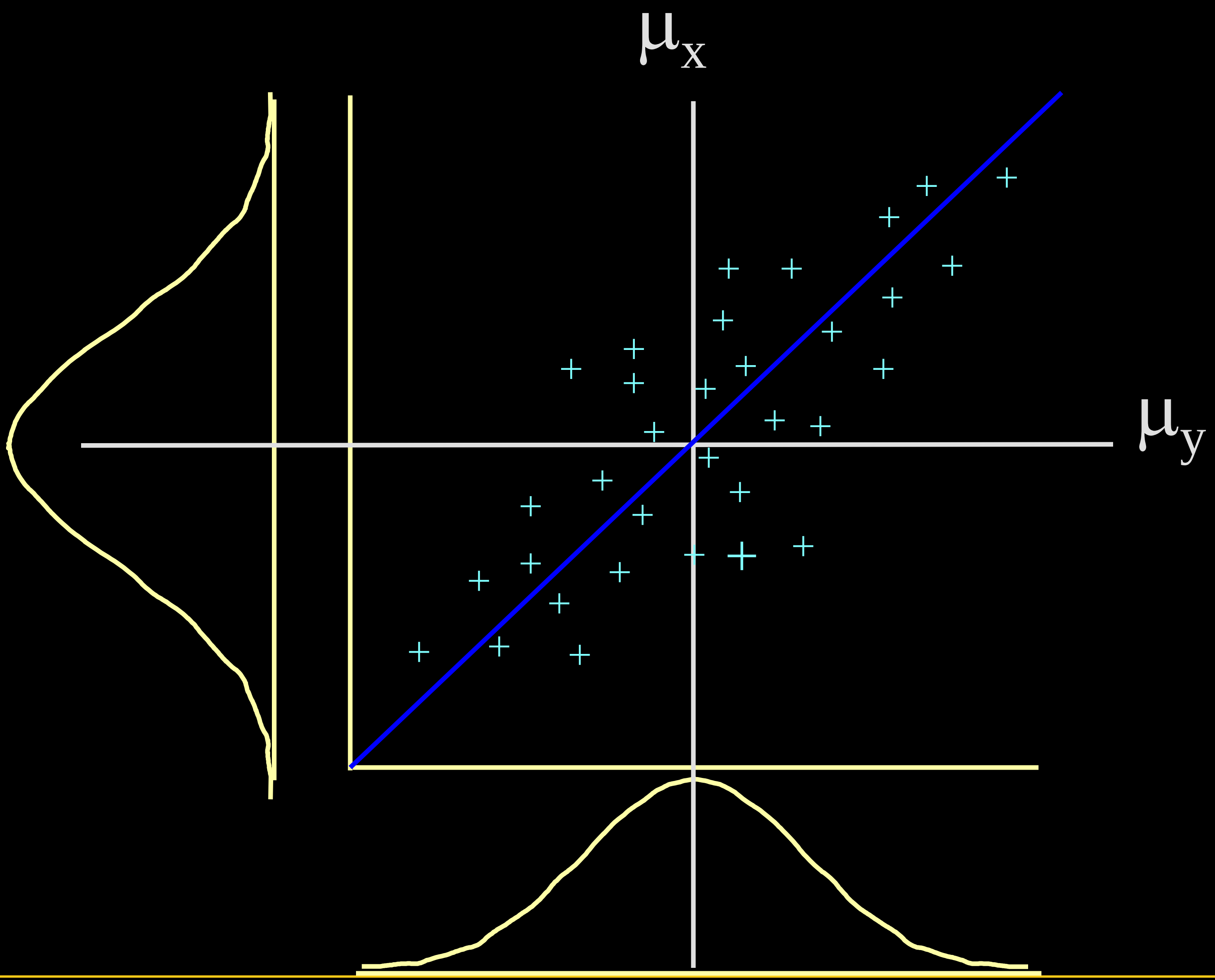
Variance: Average squared deviation

Normal distribution

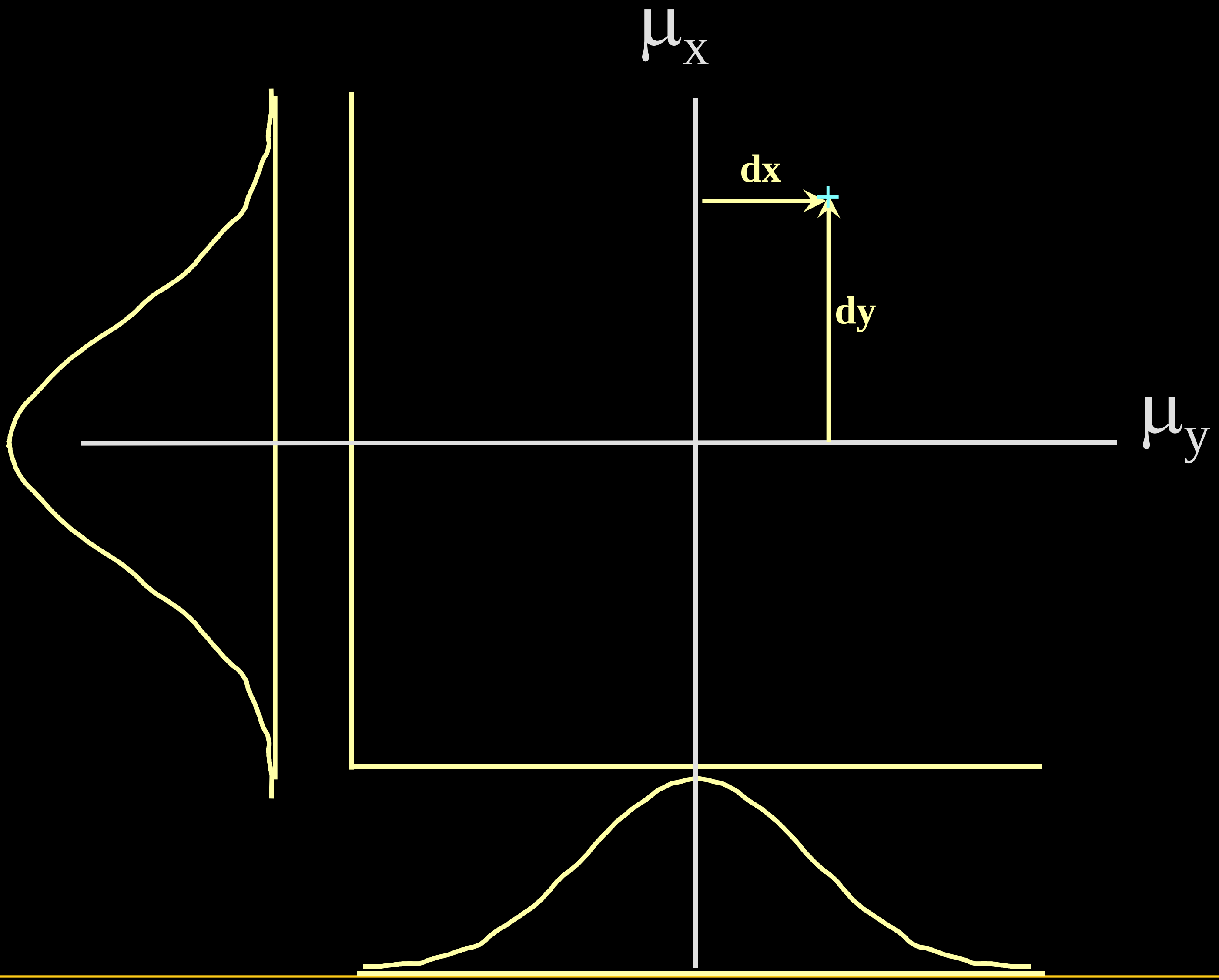


$$\text{Variance} = \Sigma\, d_i^2/N$$

Deviations in two dimensions



Deviations in two dimensions: dx x dy



Covariance

- Measure of association between two variables
- Closely related to variance
- Useful to partition variance
  - "Analysis of Variance" term coined by Fisher

Variance covariance matrix

Univariate Twin/Sib Data

$$\begin{bmatrix} \text{Var(Twin1)} & \text{Cov(Twin1,Twin2)} \\ \text{Cov(Twin2,Twin1)} & \text{Var(Twin2)} \end{bmatrix}$$

Suitable for modeling when no missing data
Good conceptual perspective

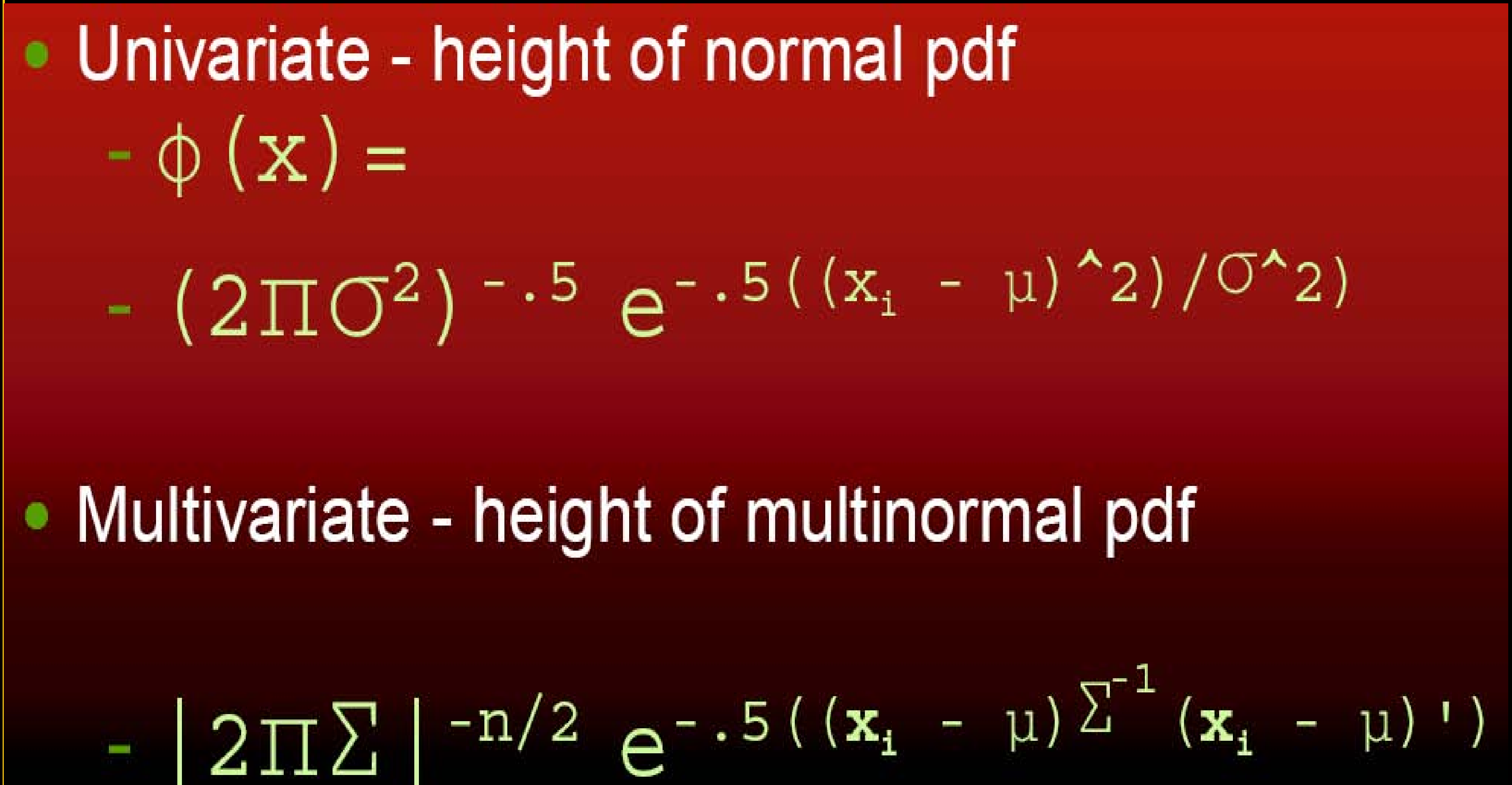
Maximum Likelihood Estimates: Nice Properties

1. Asymptotically unbiased
   - Large sample estimate of p -> population value
2. Minimum variance “Efficient”
   - Smallest variance of all estimates with property 1
3. Functionally invariant
   - If g(a) is one-to-one function of parameter a
   - and MLE (a) = a*
   - then MLE g(a) = g(a*)

- See http://wikipedia.org

Full Information Maximum Likelihood (FIML)

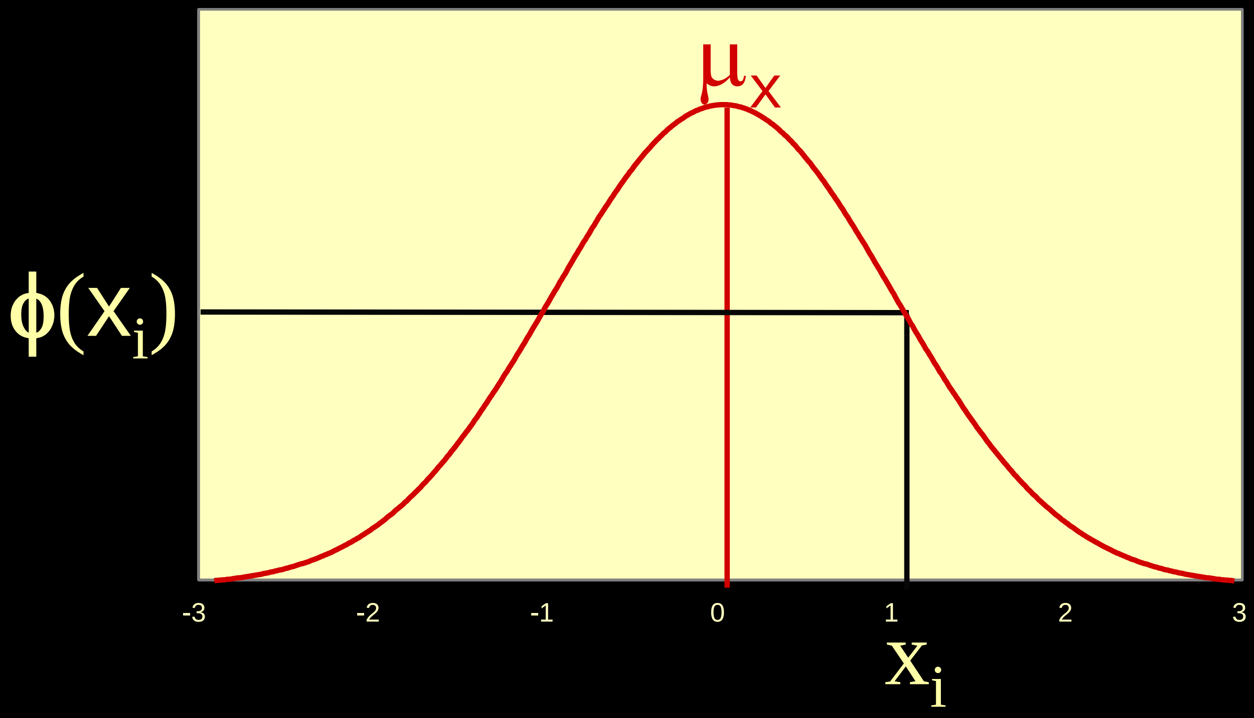
Calculate height of curve for each raw data vector



- Univariate - height of normal pdf
  - $\phi(x) =$
  - $(2\Pi\sigma^2)^{-.5}\, e^{-.5((x_i - \mu)\hat{}2)/\sigma\hat{}2)}$
- Multivariate - height of multinormal pdf
  - $|2\Pi\Sigma|^{-n/2}\, e^{-.5((\mathbf{x}_i - \mu)\Sigma^{-1}(\mathbf{x}_i - \mu)')}$

Height of normal curve: $\mu_x = 0$

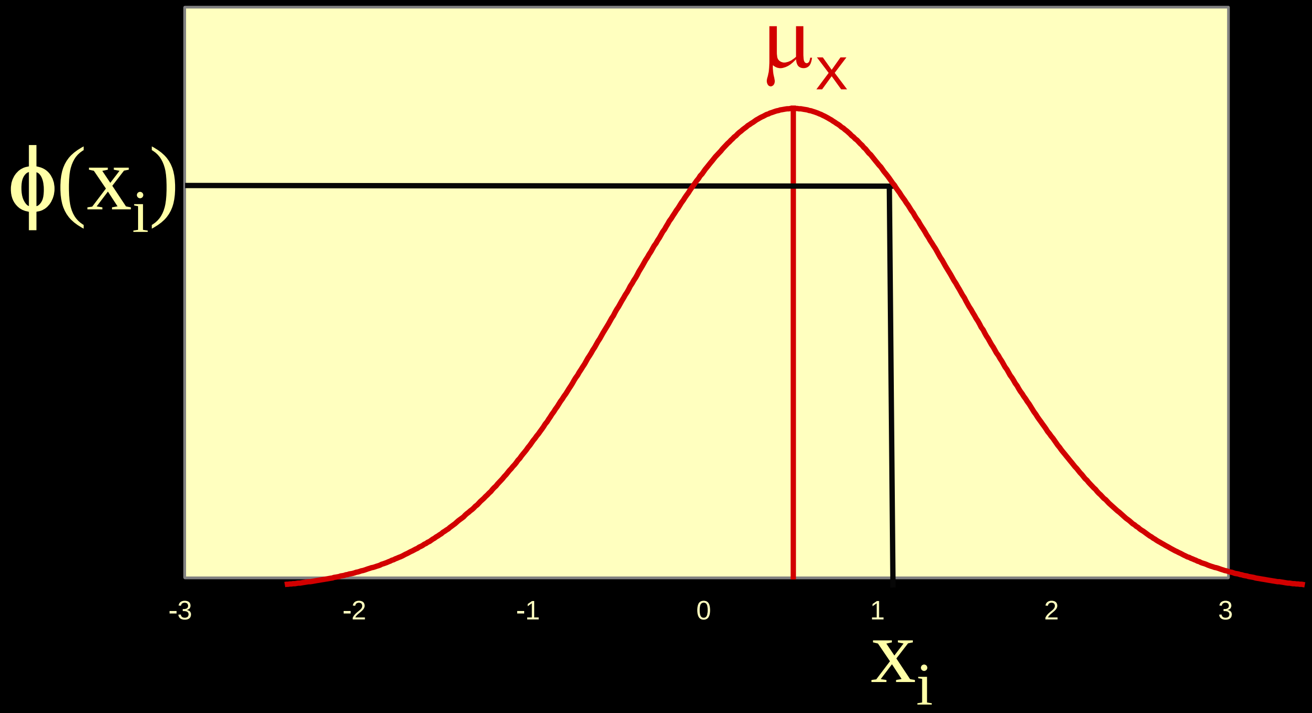
Probability density function



$\phi(x_i)$ is the likelihood of data point x_i for particular mean & variance estimates

Height of normal curve at xi: $\mu_x = .5$

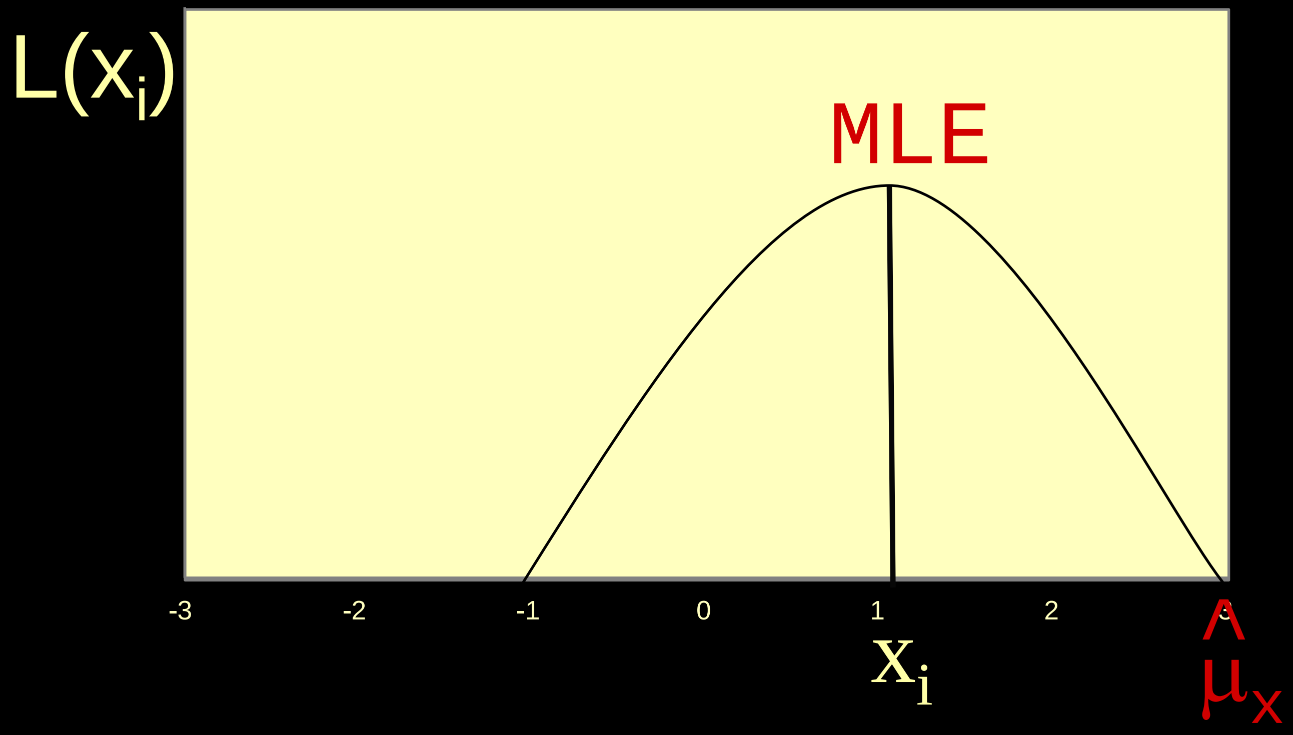
Function of *mean*



Likelihood of data point x_i *increases* as μ_x approaches x_i

Likelihood of x_i as a function of μ

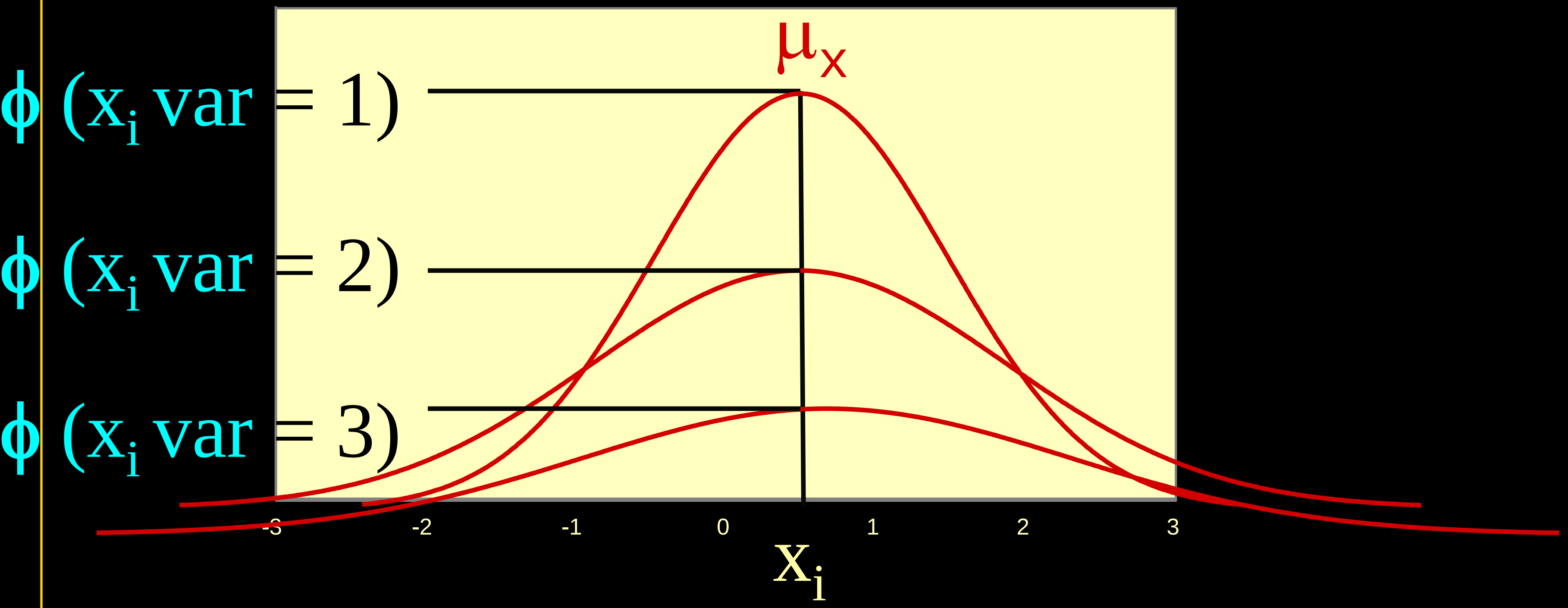
Likelihood function



$L(x_i)$ is the likelihood of data point x_i for particular mean & variance estimates

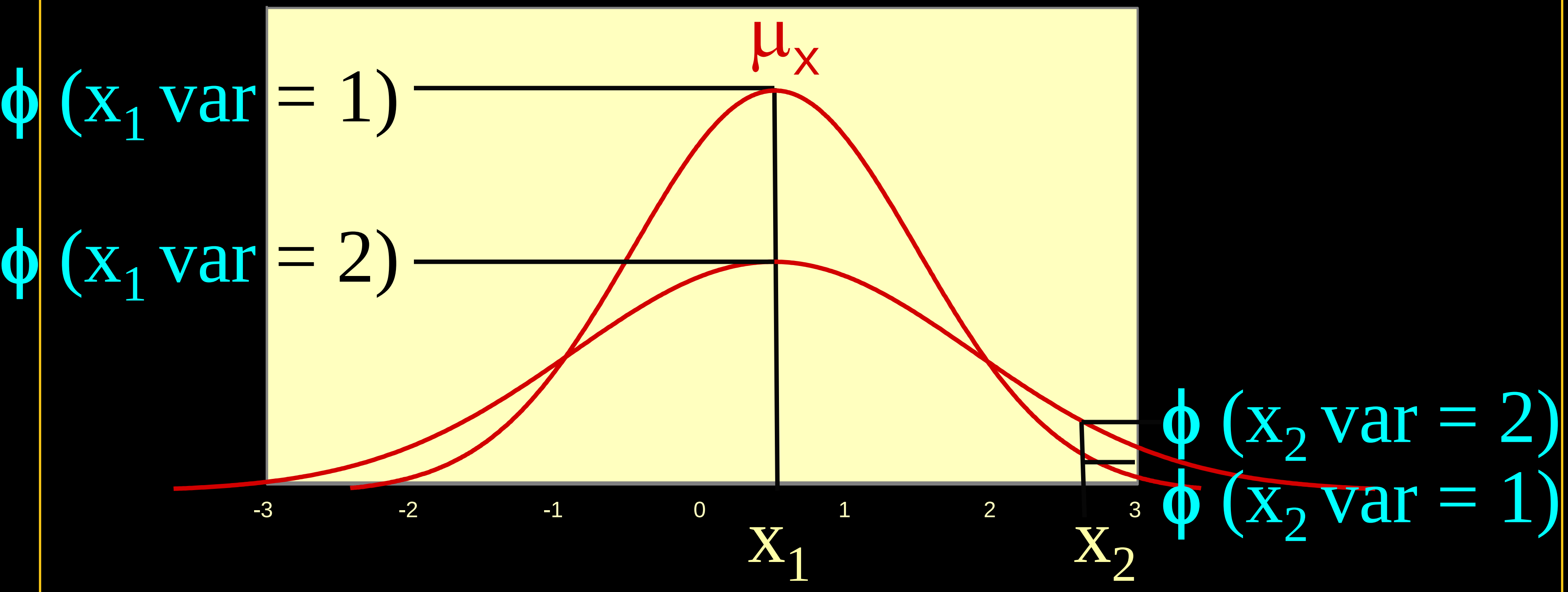
Height of normal curve at x1

Function of *variance*



Likelihood of data point x_i *changes* as variance of distribution changes

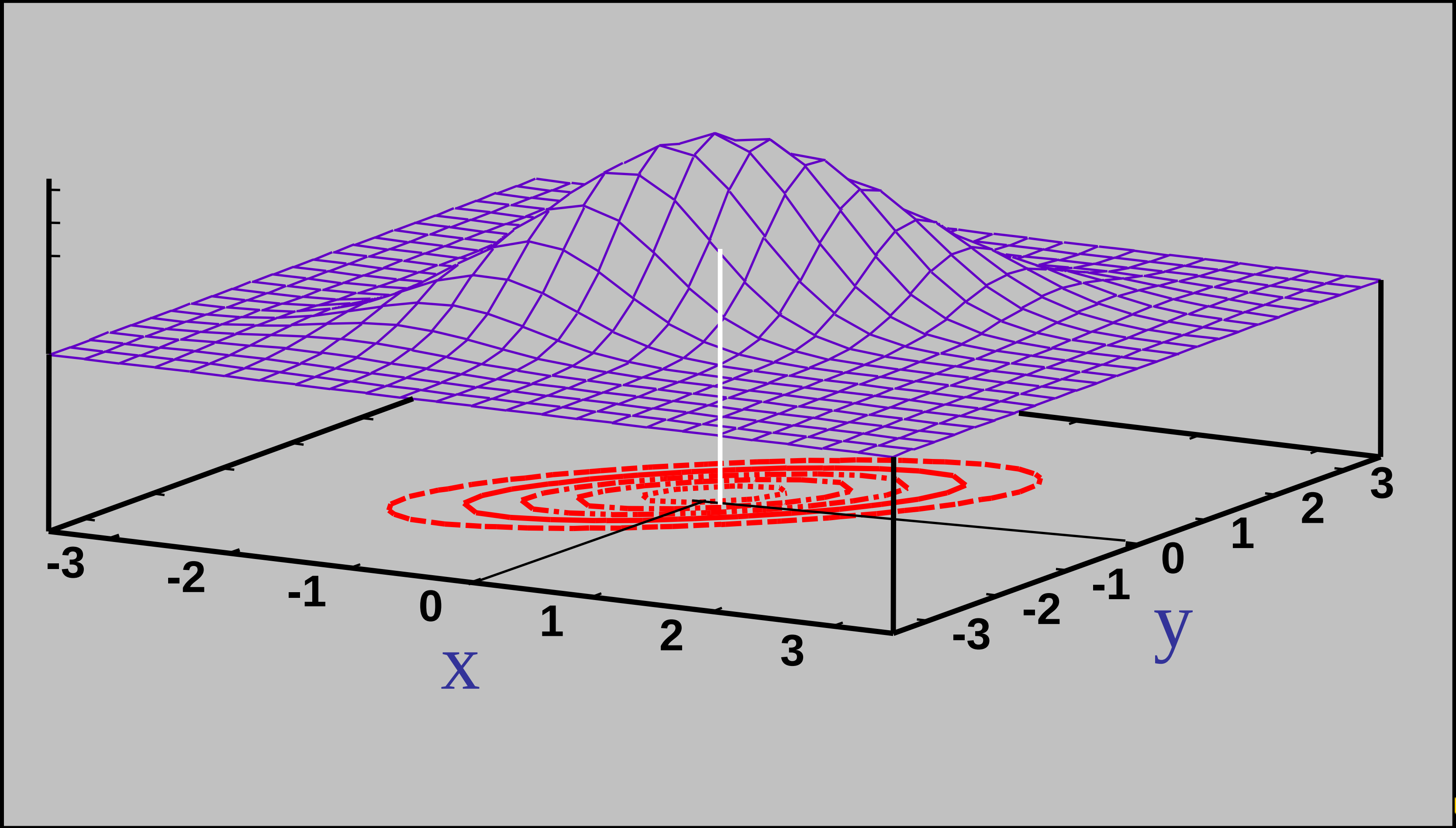
Height of normal curve at x1 and x2



x_1 has higher likelihood with var=1 whereas x_2 has higher likelihood with var=2

Height of bivariate normal density function

Likelihood varies as f(μ_1, μ_2, σ_1, σ_2, ρ)



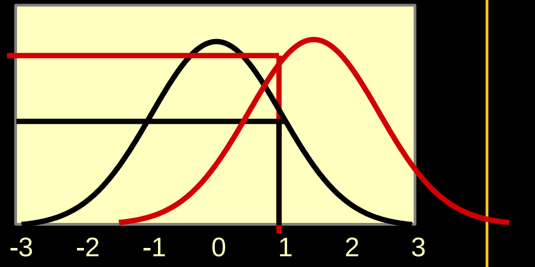
Likelihood of Independent Observations

- Chance of getting two heads
- $L(x_1 \ldots x_n)$ = Product $(L(x_1),\ L(x_2)\ ,\ \ldots\ L(x_n))$
- $L(x_i)$ typically < 1

- Avoid vanishing $L(x_1 \ldots x_n)$
- Computationally convenient log-likelihood
- $\ln (a * b) = \ln(a) + \ln(b)$

- Minimization more manageable than maximization
  - Minimize $-2 \ln(L)$

Likelihood Ratio Tests



- Comparison of likelihoods
- Consider *ratio* L(data,model 1) / L(data, model 2)
- ln(a/b) = ln(a) - ln(b)
- Log-likelihood lnL(data, model 1) - ln L(data, model 2)

- Useful asymptotic feature when model 2 is a submodel of model 1

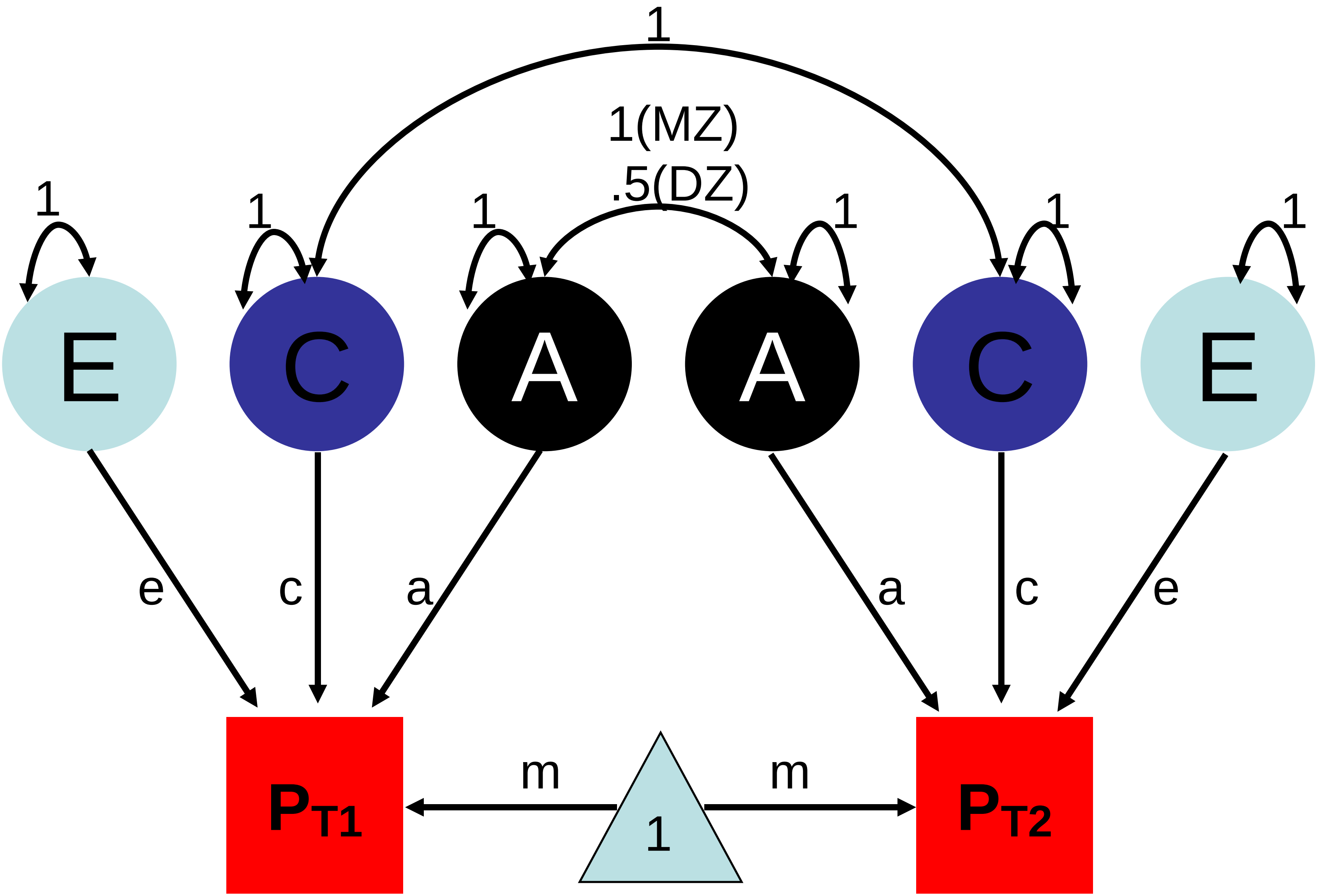
  -2 (lnL(data, model 1) - lnL(data, model 2)) ~ χ^2

  df = # parameters of model 1 - # parameters of model 2

- BEWARE of gotchas!
  - Estimates of a^2 q^2 etc. have implicit bound of zero
  - Distributed as 50:50 mixture of 0 and $\chi_1{}^2$

VCU

Two Group ACE Model for twin data



Linkage vs Association

Linkage	Association
1. Family-based	1. Families or unrelated individuals
2. Matching/ethnicity generally unimportant	2. Matching/ethnicity crucial
3. Few markers for genome coverage (300-400 STRs)	3. Many markers req for genome coverage (10^5 – 10^6 SNPs)
4. Can be weak design	4. Powerful design
5. Good for initial detection; poor for fine-mapping	5. Ok for initial detection; good for fine-mapping
6. Powerful for rare variants	6. Powerful for common variants; rare variants generally impossible

Identity by Descent (IBD)

Number of alleles shared IBD at a locus,
parents AB and CD: Three subgroups of sibpairs

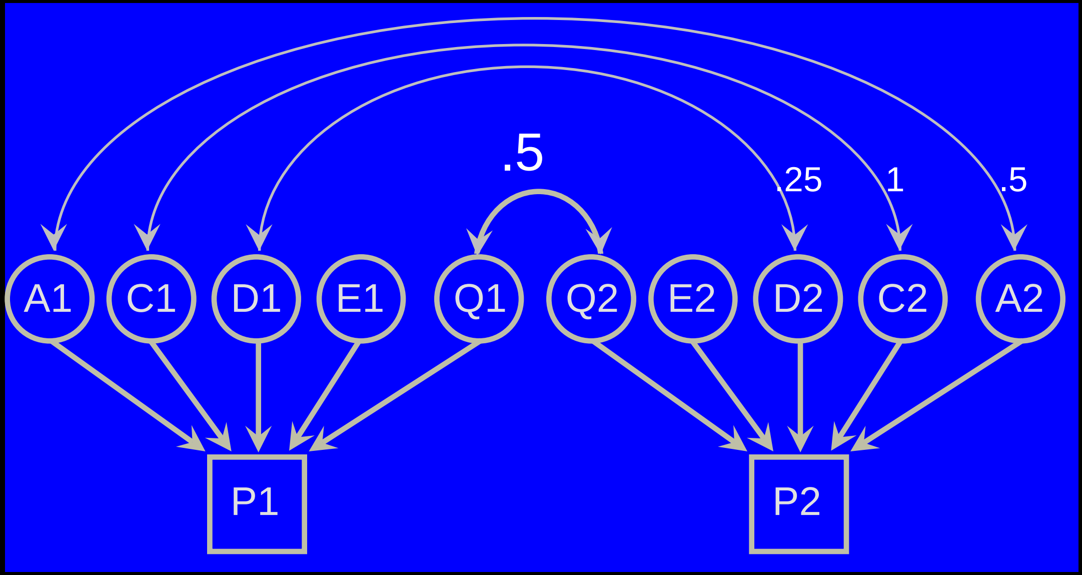
	AC	AD	BC	BD
AC	2	1	1	0
AD	1	2	0	1
BC	1	0	2	1
BD	0	1	1	2

Partitioned Twin Analysis

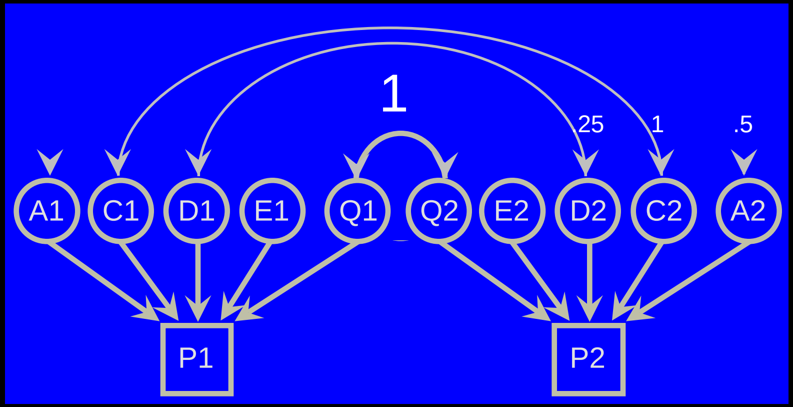
- Nance & Neale (1989) Behav Genet 19:1
  - Separate DZ pairs into subgroups
    - IBD=0 IBD=1 IBD=2
  - Correlate Q with 0 .5 and 1 coefficients
  - Compute statistical power

Partitioned Twin Analysis: Three DZ groups

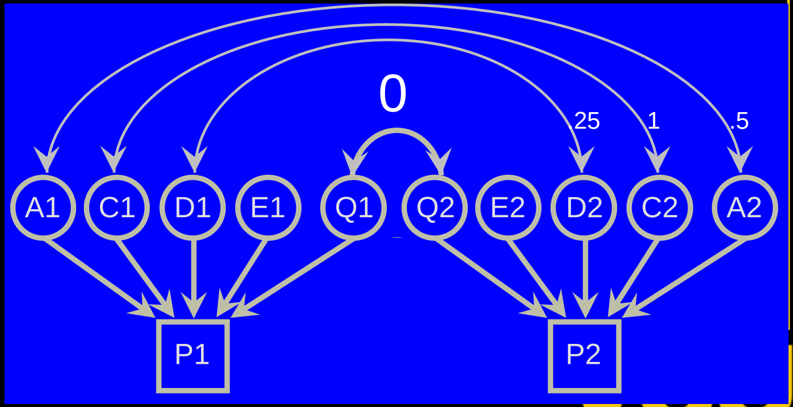
IBD=1 group



IBD=2 group



IBD=0 group



Problem 1 with Partitioned Twin analysis: Low Power

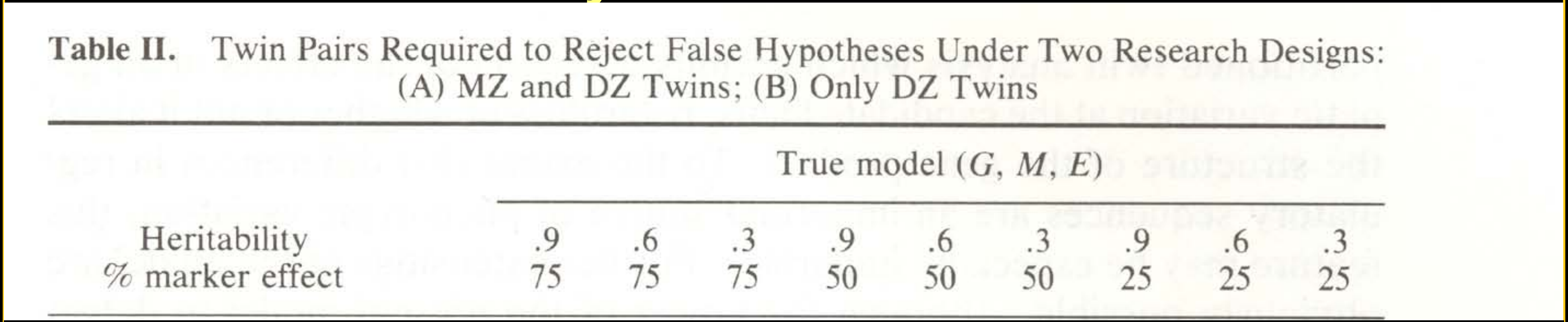
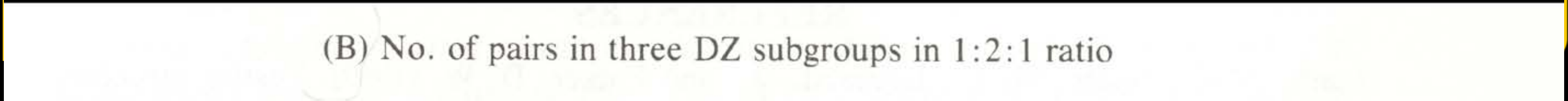


Table II. Twin Pairs Required to Reject False Hypotheses Under Two Research Designs: (A) MZ and DZ Twins; (B) Only DZ Twins

	True model (G, M, E)								
Heritability	.9	.6	.3	.9	.6	.3	.9	.6	.3
% marker effect	75	75	75	50	50	50	25	25	25

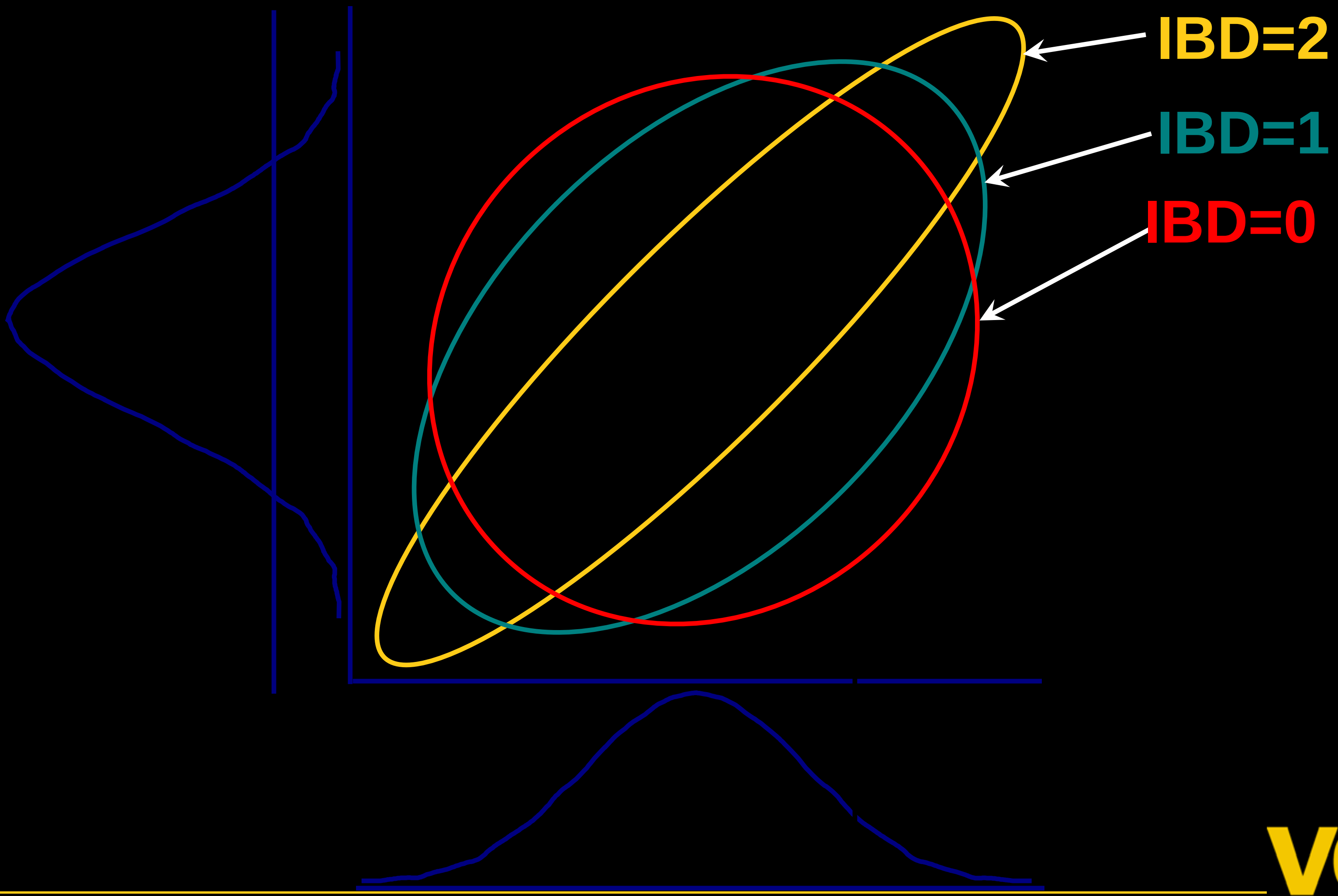


(B) No. of pairs in three DZ subgroups in 1:2:1 ratio

Problem 2: IBD is not known with certainty

- Markers may not be fully informative
  - Only so much heterozygosity in e.g., 20 allele microsatellite marker
  - Less in a SNP
  - Unlikely to have typed the exact locus we are looking for
  - Genome is big!

IBD pairs vary in similarity

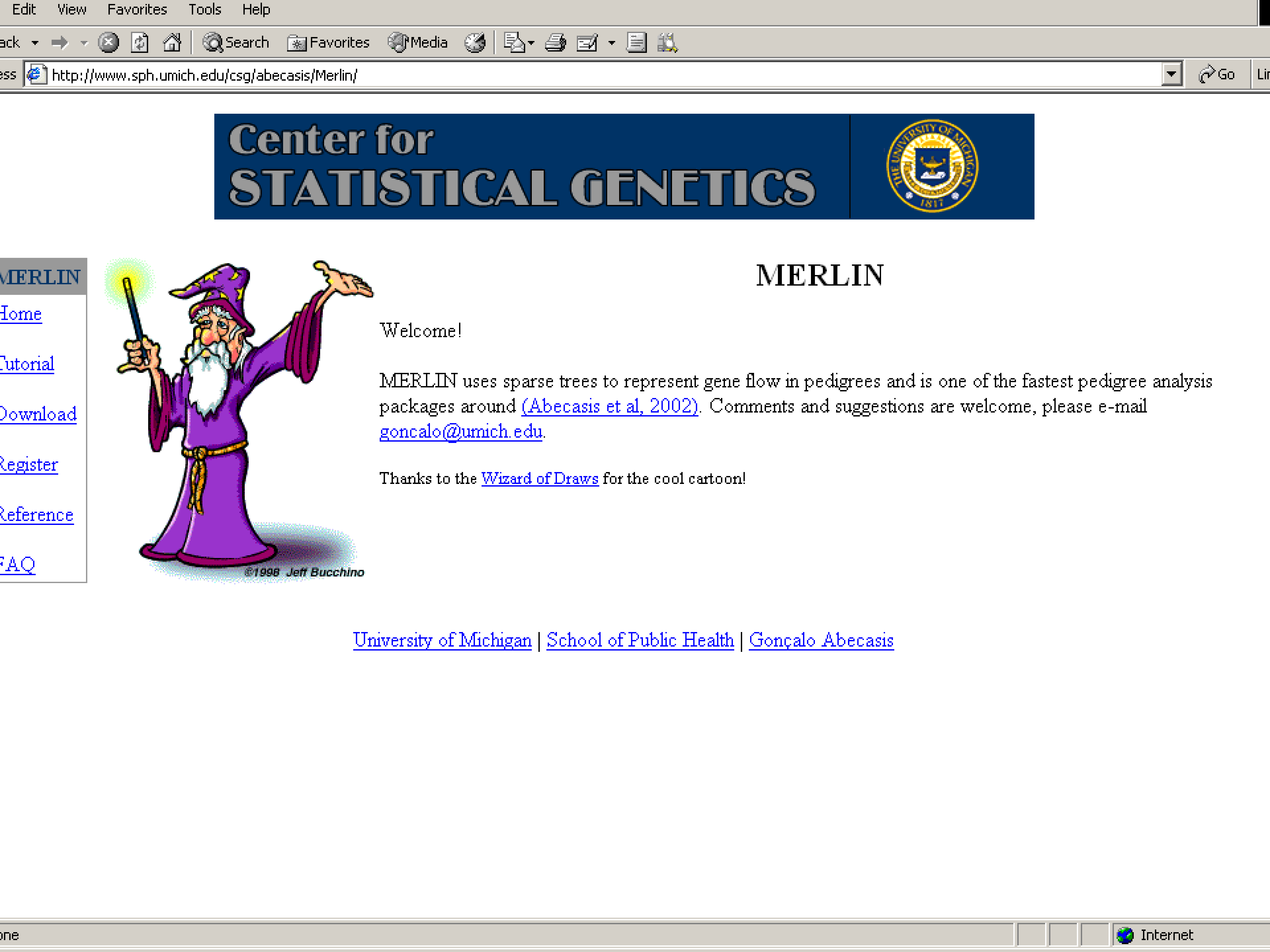


Improving Power for Linkage

- Increase marker density (yaay SNP chips)
- Change design
  - Families
  - Larger Sibships
  - Selected samples
- Multivariate data
- More heritable traits with less error

Problem 2: IBD is not known with certainty

- Markers may not be fully informative
  - Only so much heterozygosity in e.g., 20 allele microsatellite marker
  - Less in a SNP
  - Unlikely to have typed the locus that causes variation
  - Genome is big!
  - *The Universe is Big. Really big. It may seem like a long way to the corner chemist, but compared to the Universe, that's peanuts.* - D. Adams



Edit View Favorites Tools Help
ack Search Favorites Media
ess http://www.sph.umich.edu/csg/abecasis/Merlin/
Go
Lin
Center for STATISTICAL GENETICS
THE UNIVERSITY OF MICHIGAN 1817
MERLIN
Home
Tutorial
Download
Register
Reference
FAQ
©1998 Jeff Bucchino
one
Internet

MERLIN

Welcome!

MERLIN uses sparse trees to represent gene flow in pedigrees and is one of the fastest pedigree analysis packages around (Abecasis et al, 2002). Comments and suggestions are welcome, please e-mail goncalo@umich.edu.

Thanks to the Wizard of Draws for the cool cartoon!

University of Michigan | School of Public Health | Gonçalo Abecasis

Using Merlin/Genehunter etc

- Several Faculty experts
  - Goncalo Abecasis
  - Sarah Medland
  - Stacey Cherny
- Possible to use Merlin via Mx GUI

“Pi-hat” approach

1 Pick a putative QTL location

2 Compute p(IBD=0) p(IBD=1) p(IBD=2) given marker data [Use Mapmaker/sibs or Merlin]

3 Compute $\hat{\pi}_i$ = p(IBD=2) + .5p(IBD=1)

4 Fit model

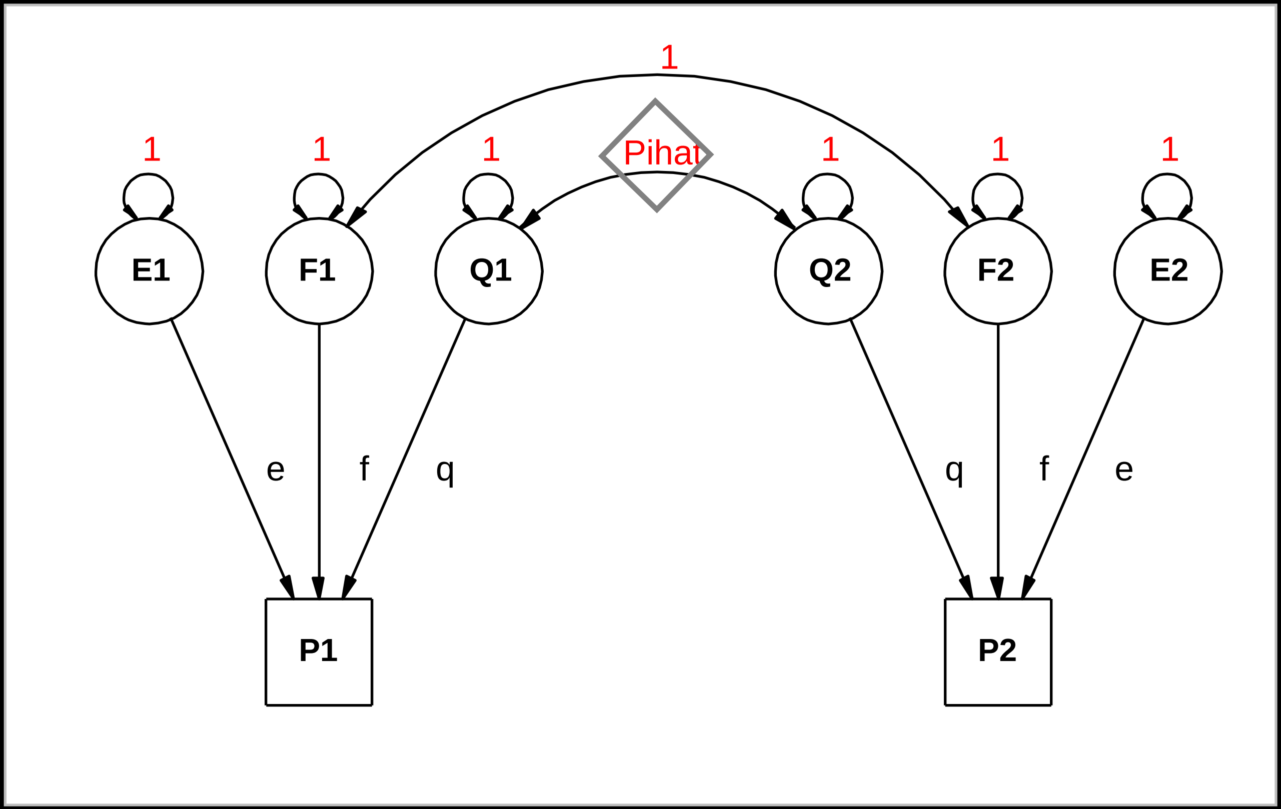
Repeat 1-4 as necessary for different locations across genome

Elston & Stewart

VCU

Basic Linkage (QTL) Model

$\pi_i = p(\mathrm{IB}\hat{\mathrm{D}}_i=2) + .5\,p(\mathrm{IBD}_i=1)$ *individual-level*



Q: QTL Additive Genetic F: Family Environment E: Random Environment

3 estimated parameters: q, f and e Every sibship may have different model

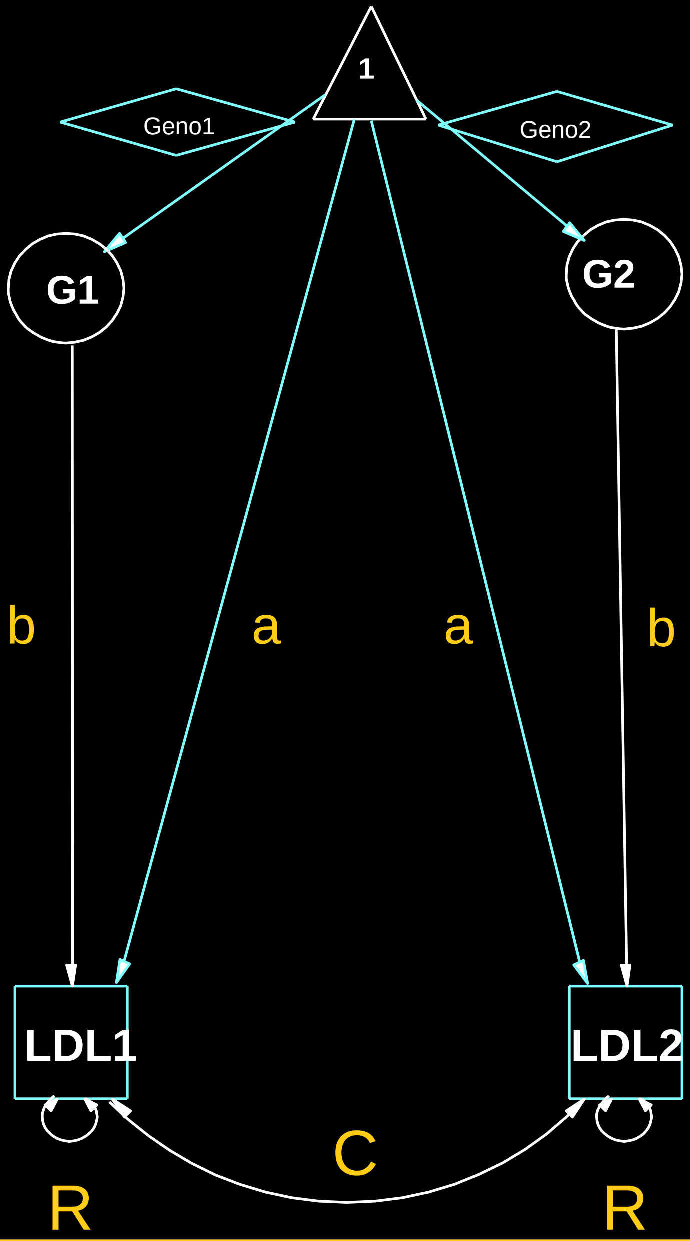
Association Model

$LDL1_i = a + b\ Geno1_i$

$Var(LDL_i) = R$

$Cov(LDL_1, LDL_2)$
$= C$

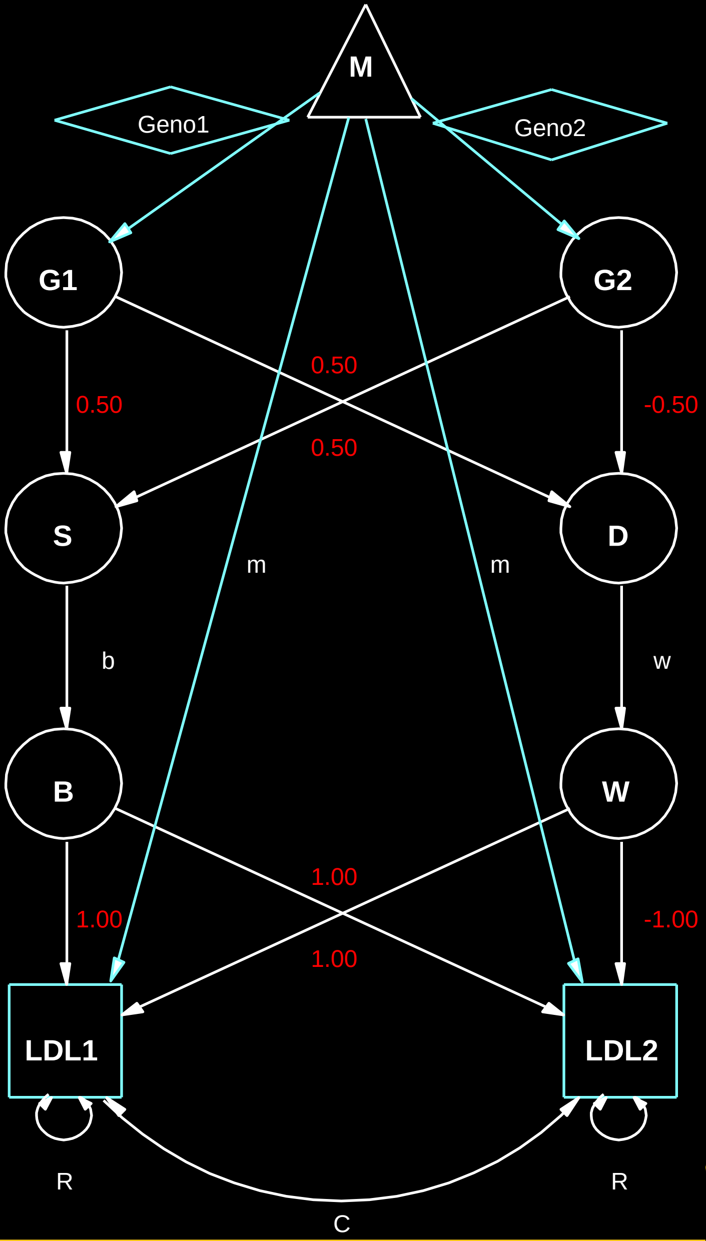
C may be f(π_i) in joint linkage & association



Between/Within Fulker Association Model

Model for the means

$$\begin{aligned} LDL1_i = {} & .5bGeno1 + \\ & .5bGeno2 + \\ & .5wGeno1 - \\ & .5wGeno2 \\ = {} & .5(\, b(Geno1+Geno2) \\ & + w(Geno1-Geno2)\,) \end{aligned}$$


M
Geno1
Geno2
G1
G2
0.50
0.50
0.50
-0.50
S
D
m
m
b
w
B
W
1.00
1.00
1.00
-1.00
LDL1
LDL2
R
R
C