Virginia Commonwealth University

Basic Statistics for Linkage and Association Studies of Quantitative Traits

Boulder Colorado Workshop March 5 2007

Overview

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

VCU

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

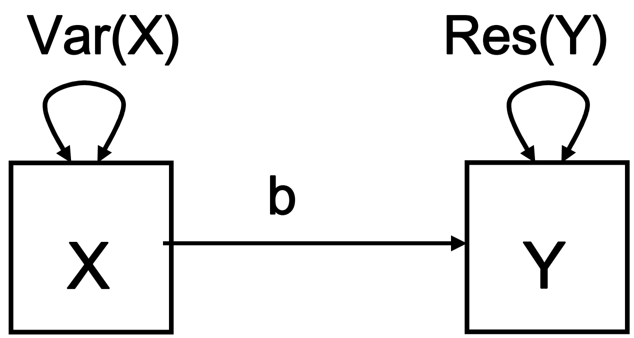
Structural Equation Model basics

- 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

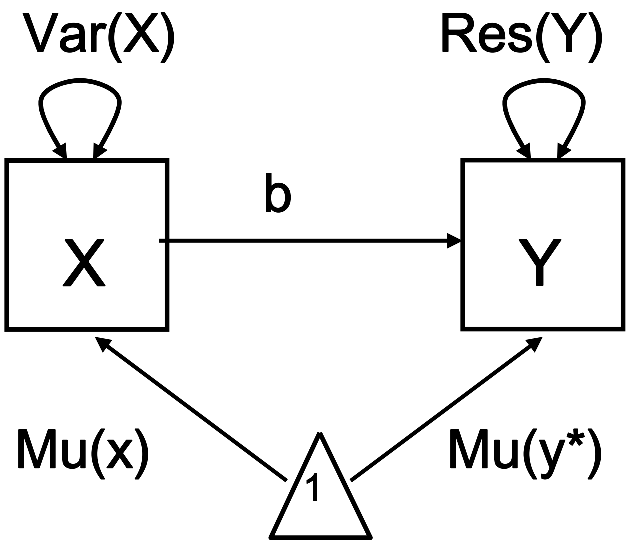
  - Triangles – constant (zero variance) for specifying means
  - Diamonds -- observed variables used as moderators (on paths)

Linear Regression Model



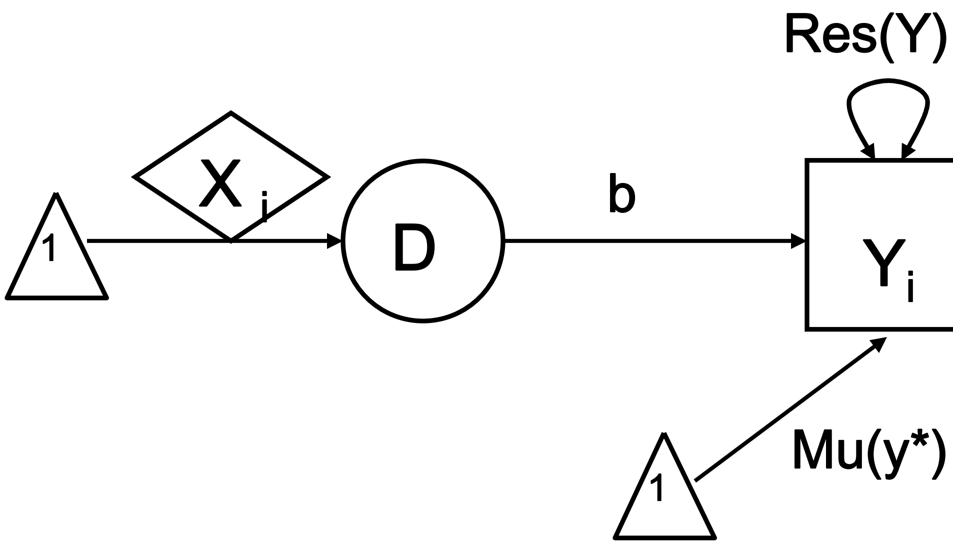
Models *covariances* only
Of historical interest

Linear Regression Model



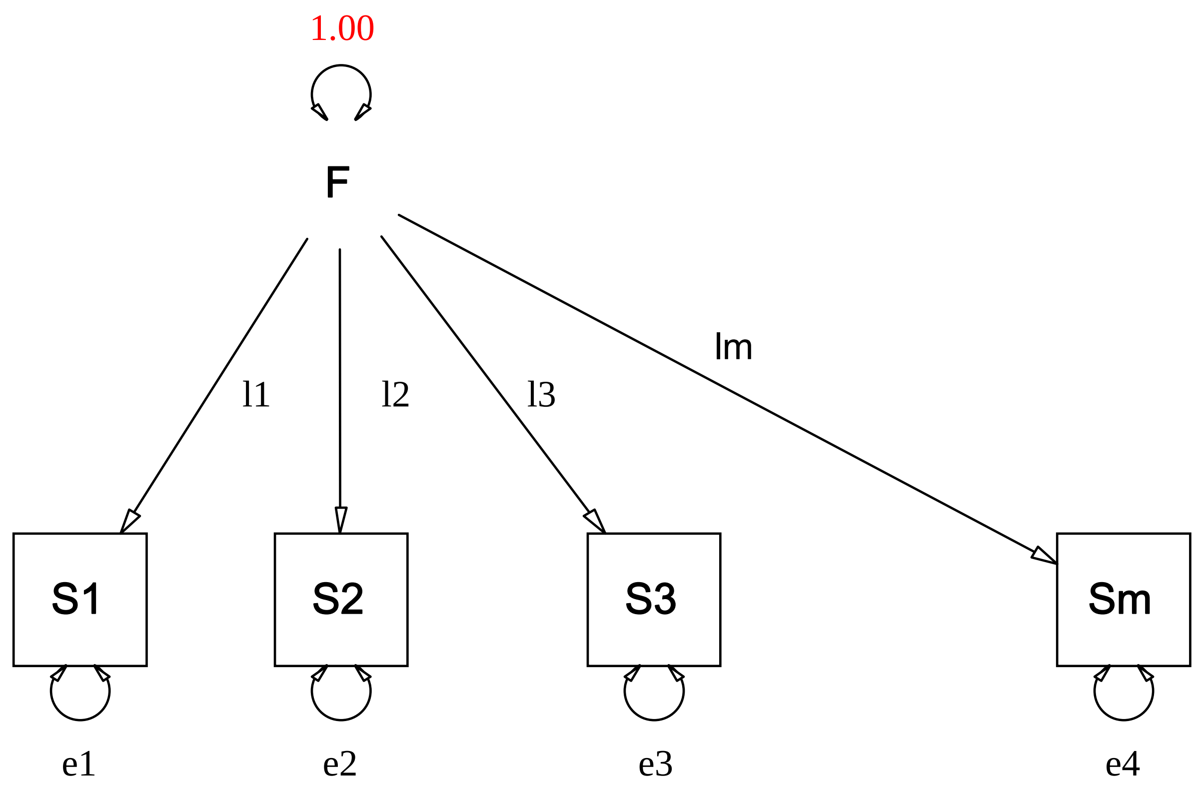
Models Means and Covariances

Linear Regression Model

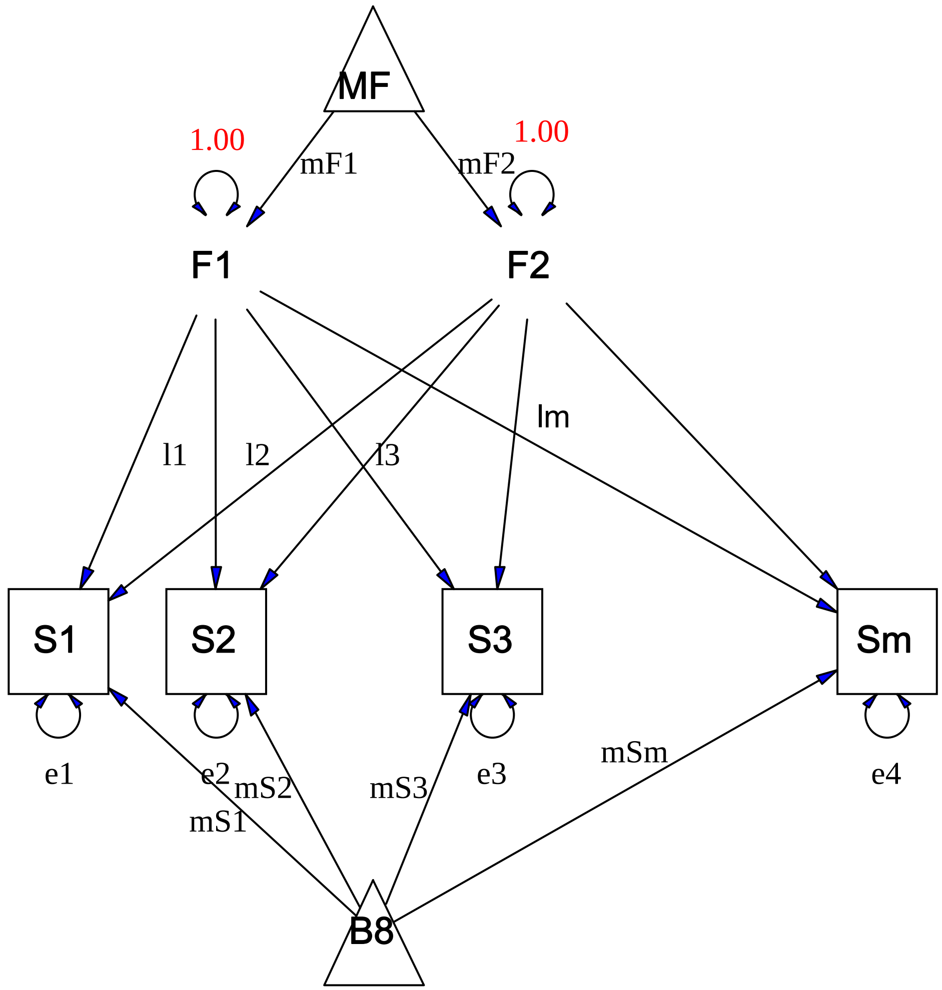


Models Mean and Covariance of Y *only*
Must have raw (individual level) data
X is a *definition* variable
Mean of Y different for every observation

Single Factor Model



Factor Model with Means



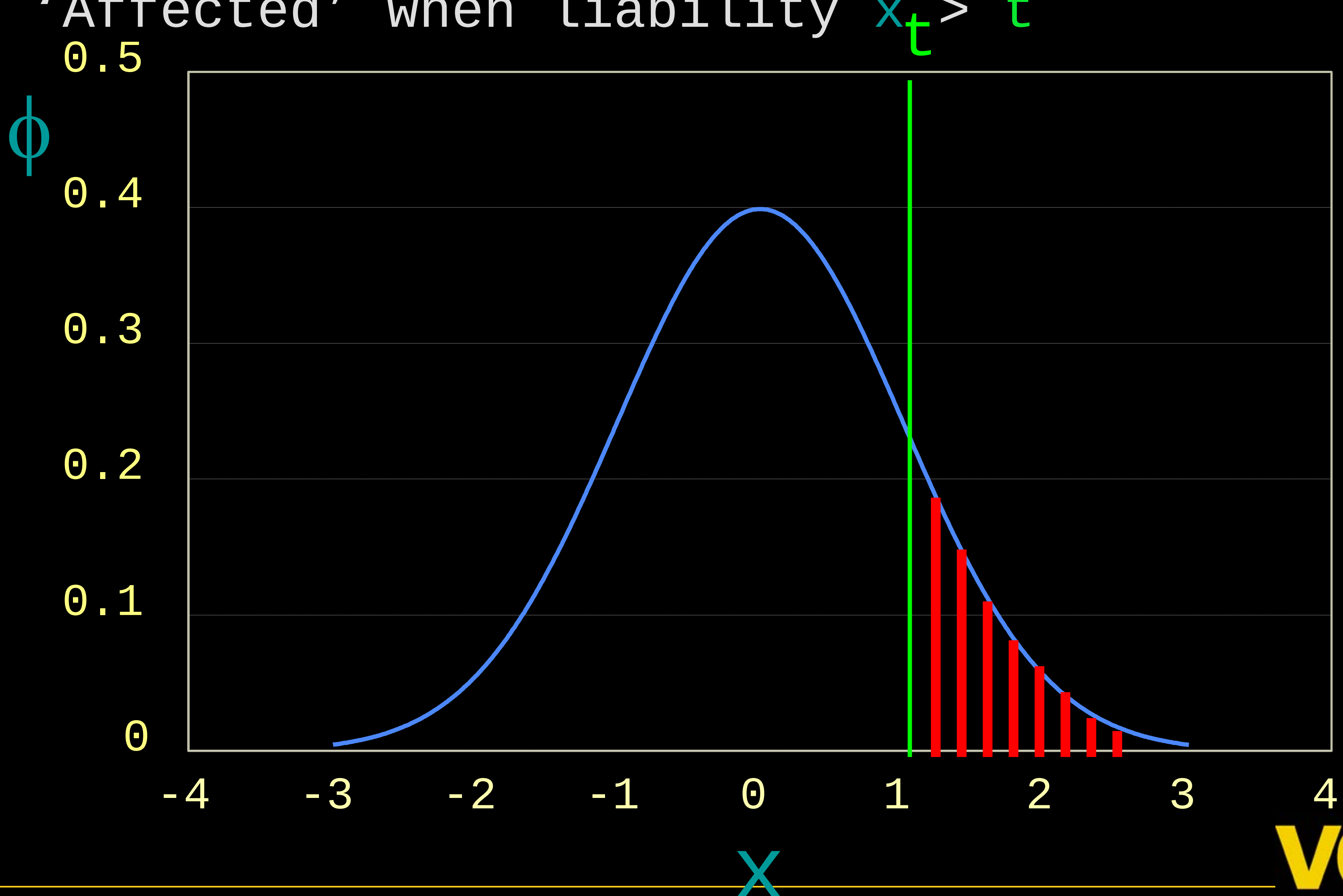
Factor model essentials

- The factor itself is typically assumed to be normally distributed: SEM
- 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 $x > t$



VCU

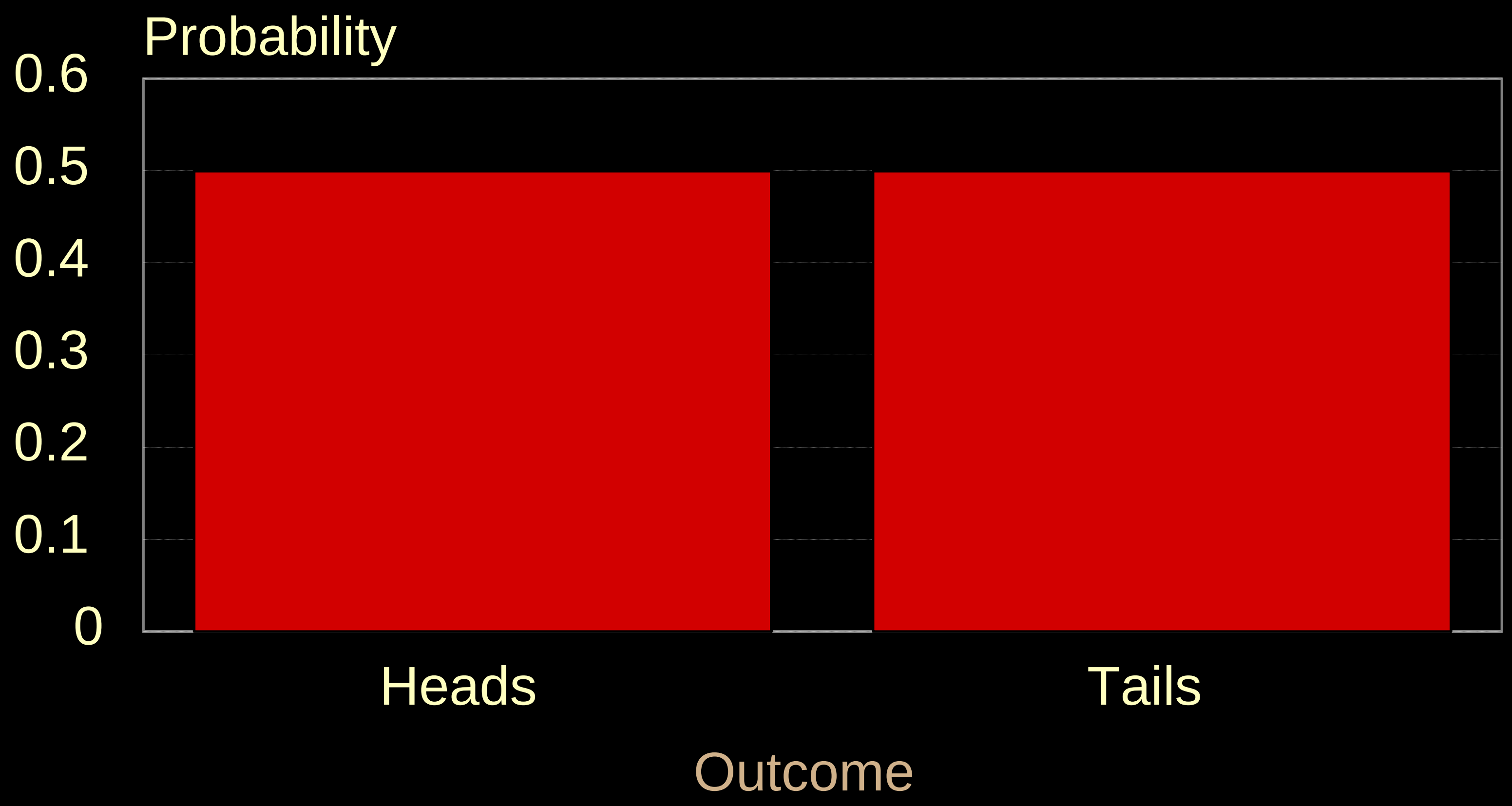
Measuring Variation

- Distribution
  - Population
  - Sample
  - Observed measures
- Probability density function ‘pdf’
  - Smoothed out histogram
  - f(x) >= 0 for all x

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

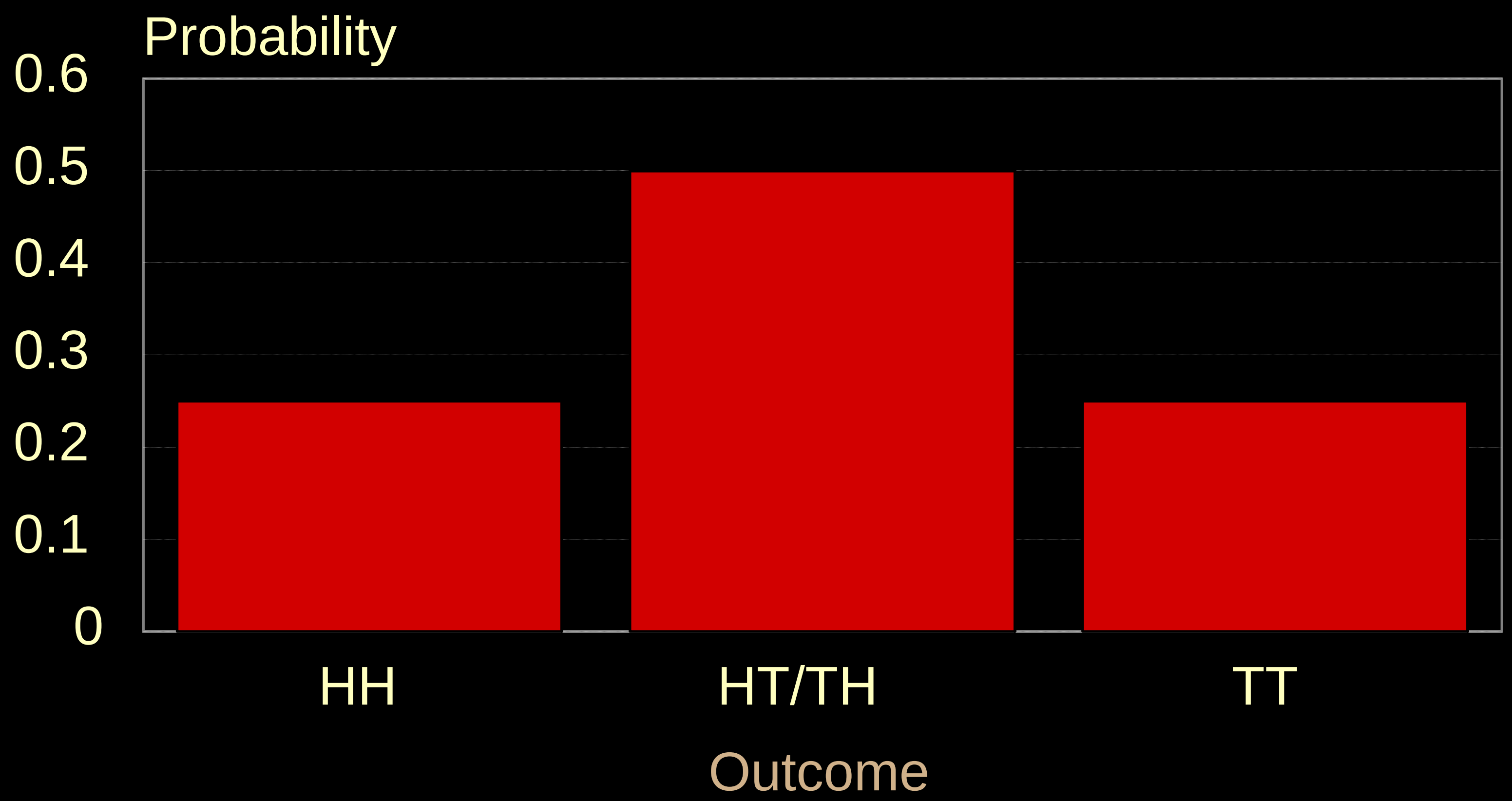
One Coin toss

2 outcomes



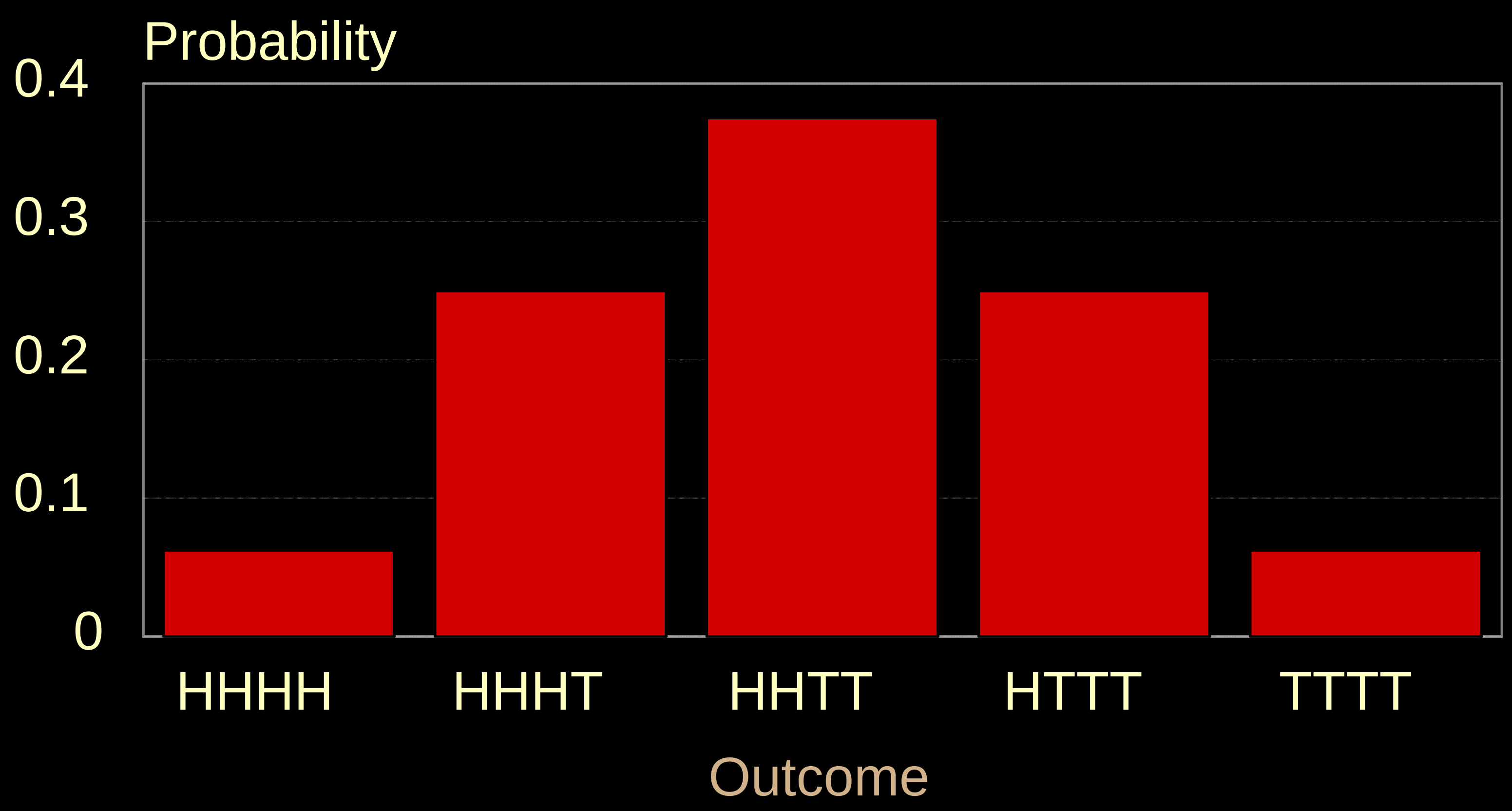
Two Coin toss

3 outcomes



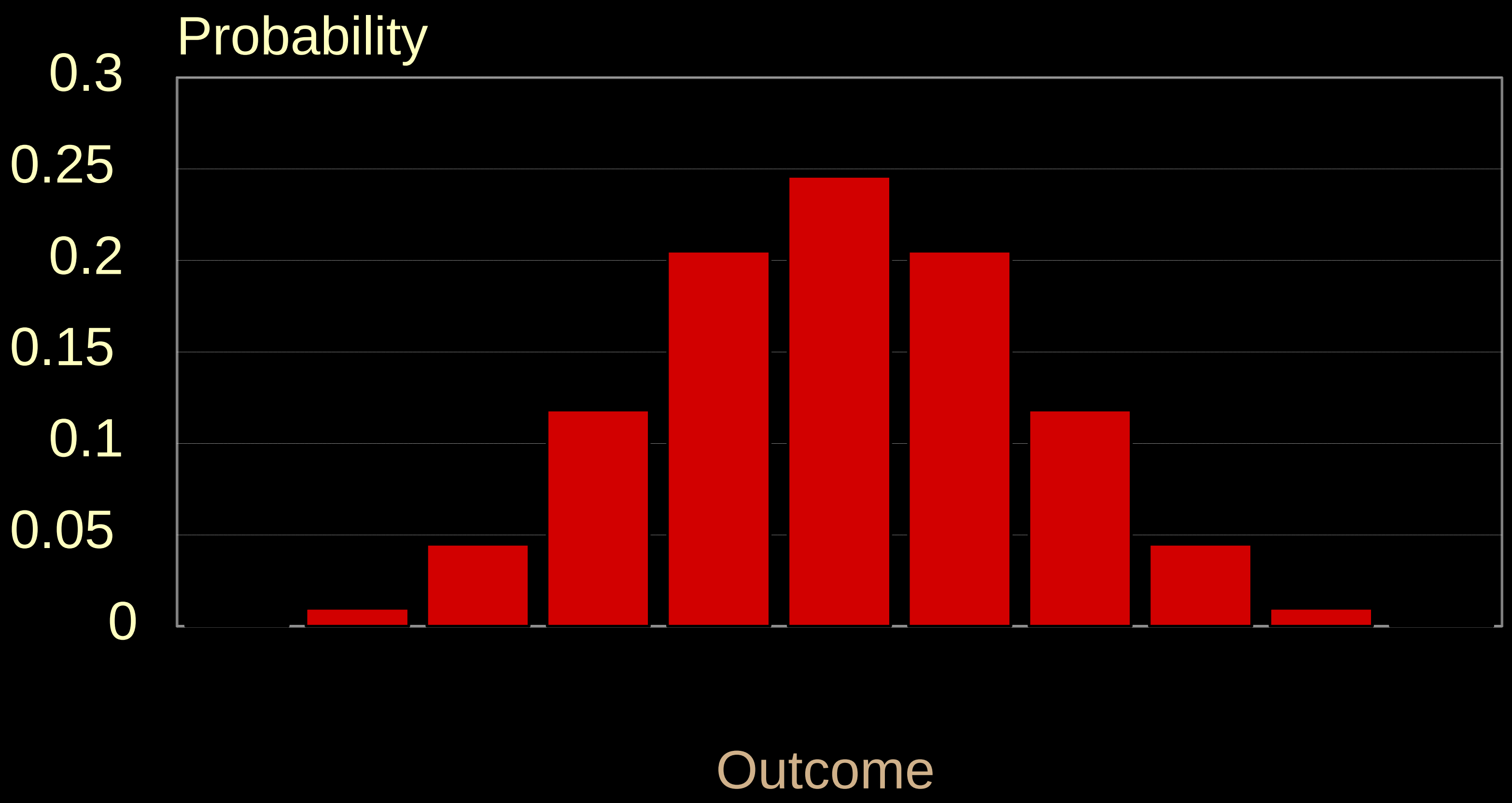
Four Coin toss

5 outcomes



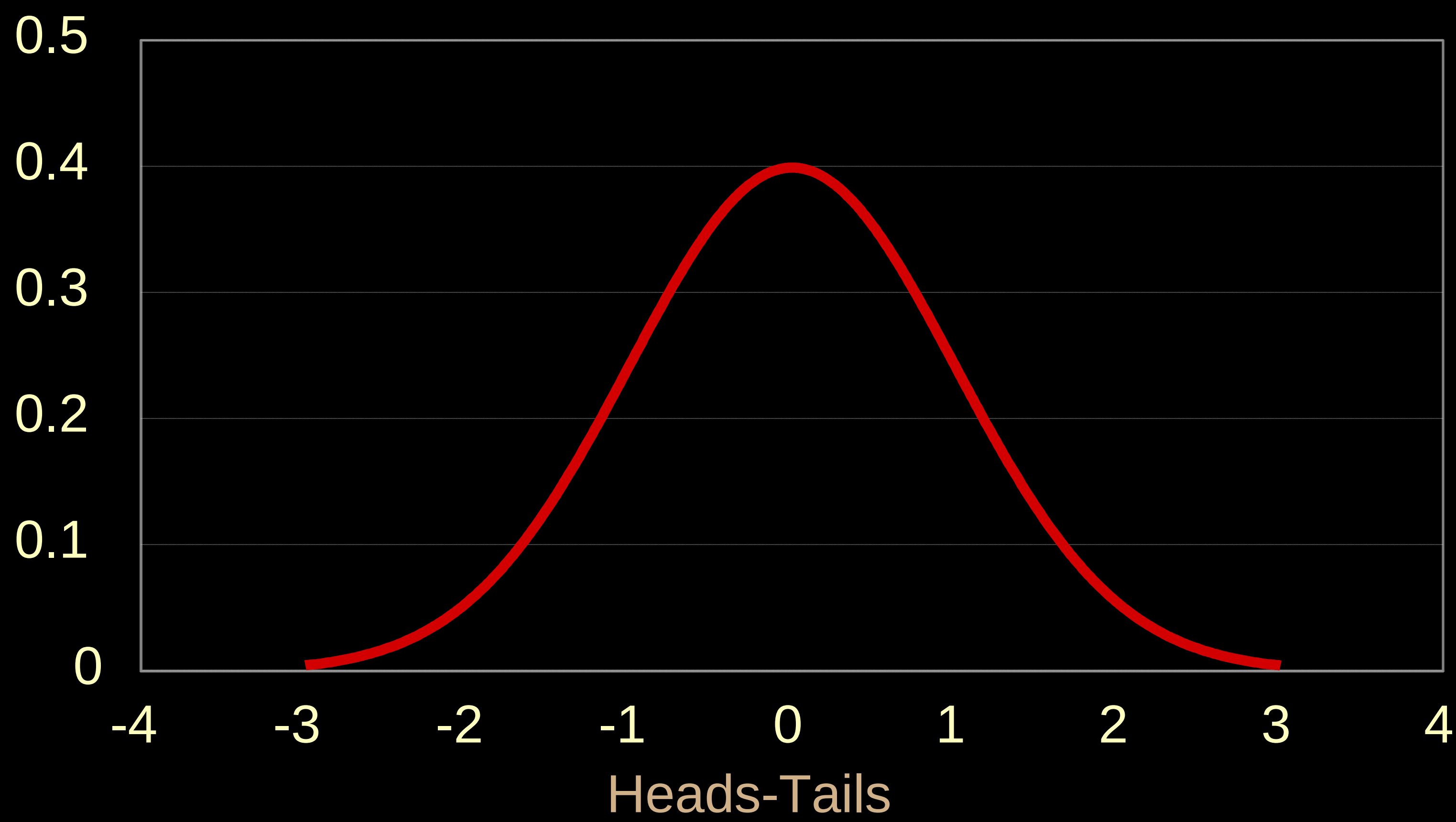
Ten Coin toss

9 outcomes



Fort Knox Toss

Infinite outcomes



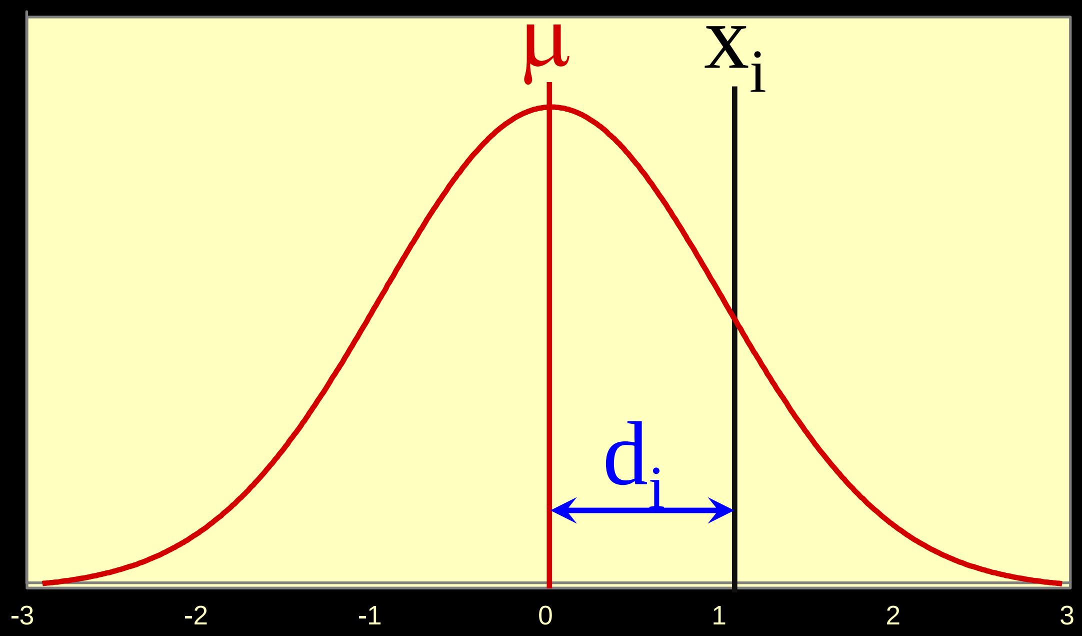
De Moivre 1733 Gauss 1827

Variance

- Measure of Spread
- Easily calculated
- Individual differences

Average squared deviation

Normal distribution



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

Measuring Variation

Weighs & Means

- Absolute differences?
- Squared differences?
- Absolute cubed?
- Squared squared?

Measuring Variation

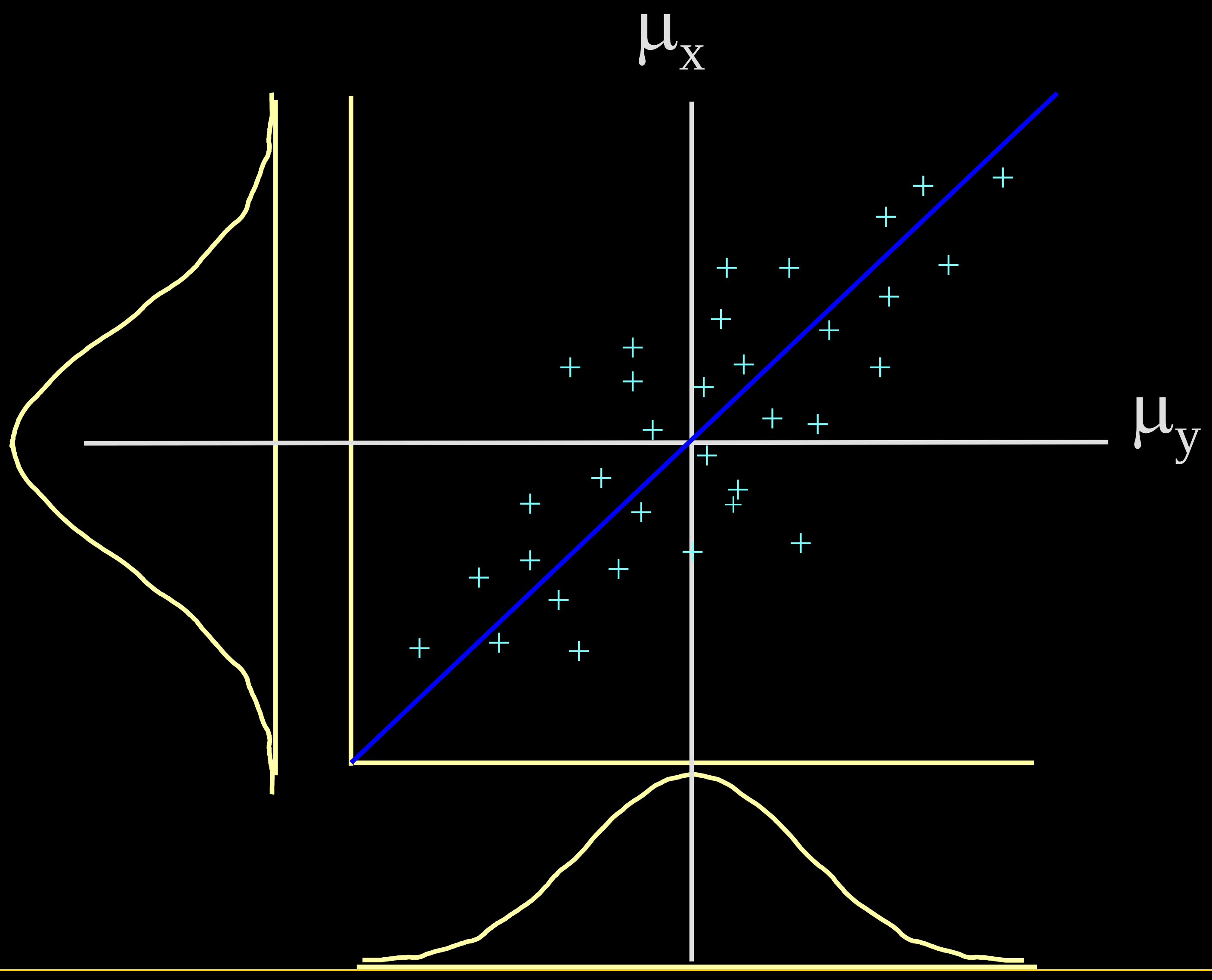
Ways & Means

- Squared differences

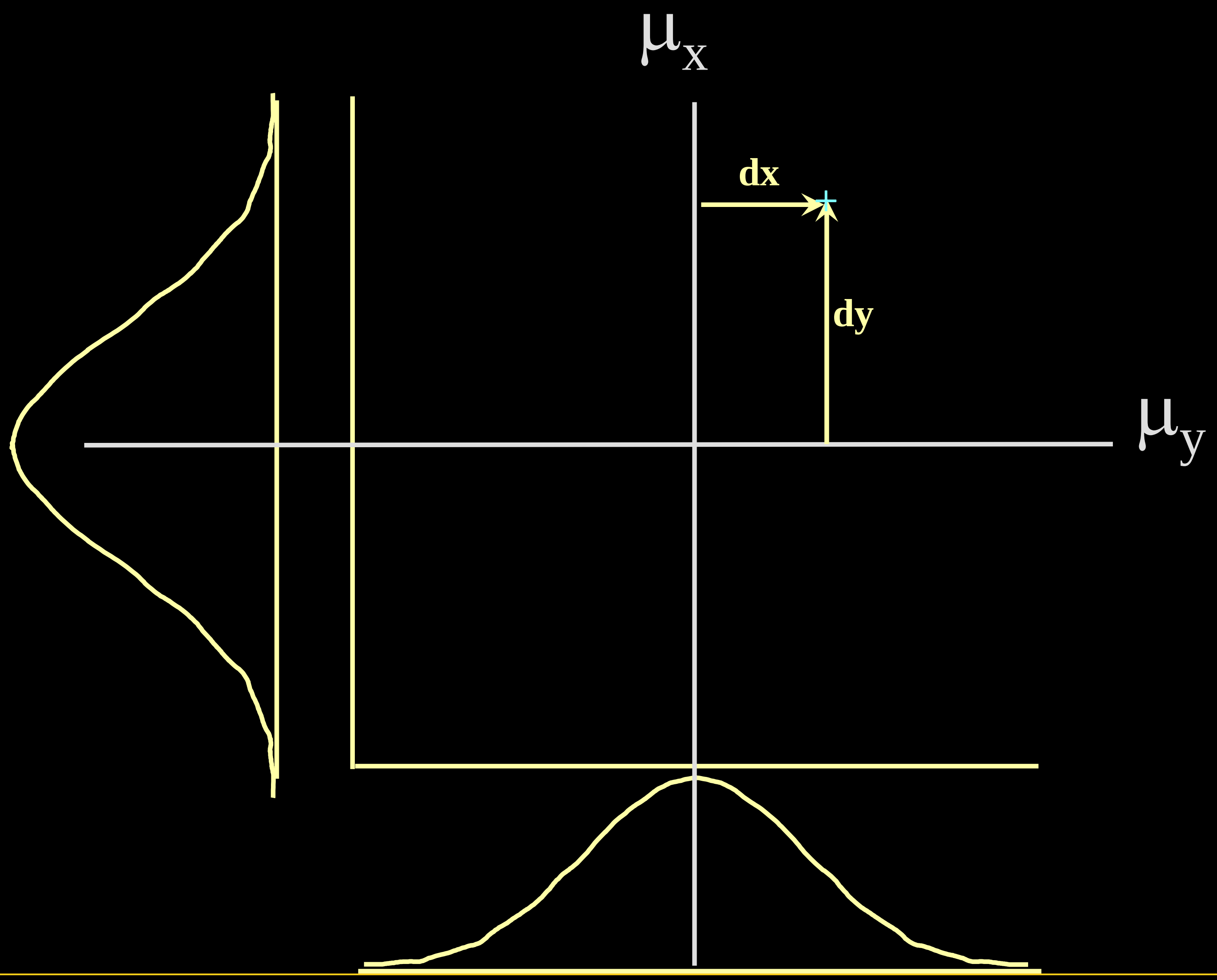
Fisher (1922) Squared has minimum variance under normal distribution

Concept of "Efficiency" emerges

Deviations in two dimensions



Deviations in two dimensions



Covariance

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

Summary

Formulae for sample statistics; i=1…N observations

$$\mu_x = (\Sigma x_i)/N$$

$$\sigma_x^2 = \Sigma\,(x_i - \mu_x)^2 / (N)$$

$$\sigma_{xy} = \Sigma\,(x_i - \mu_x)(y_i - \mu_y) / (N)$$

$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \ \sigma_x^2}}$$

Variance covariance matrix

Univariate Twin/Sib Data

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

Only suitable for complete data
Good conceptual perspective

Summary

- Means and covariances
- Basic input statistics for “Traditional SEM”
- Notion of probability density function

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

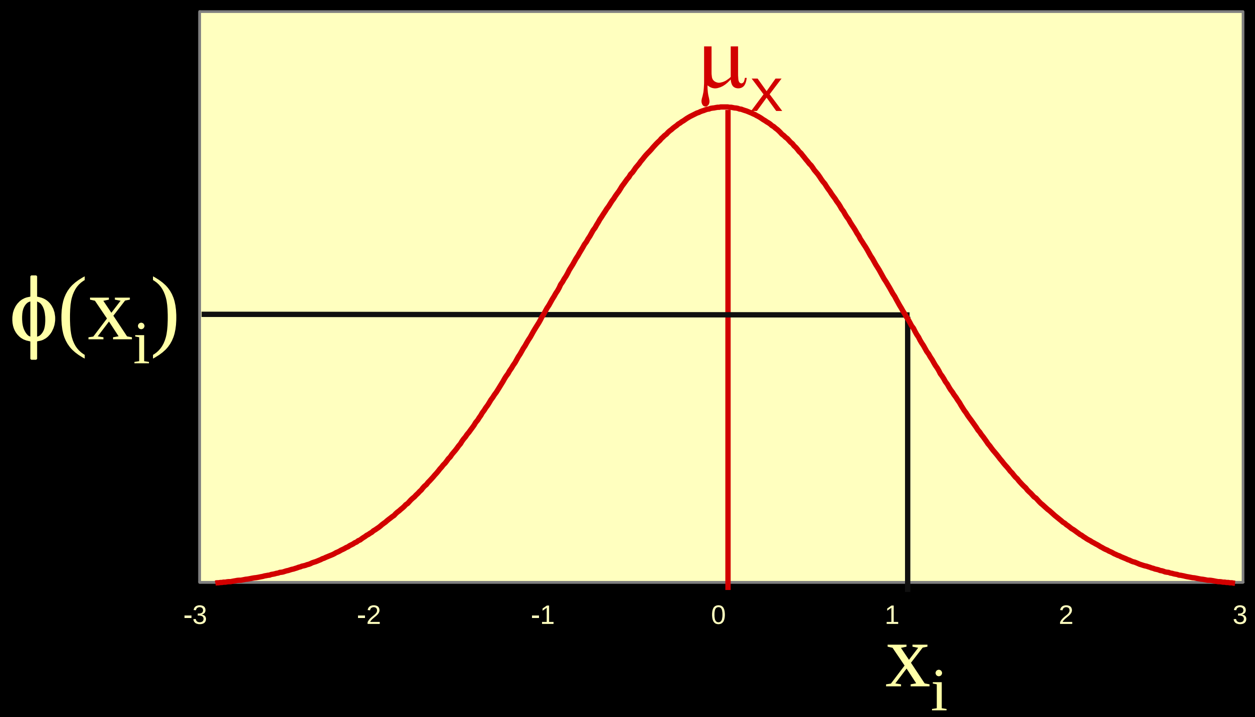
Likelihood computation

Calculate height of curve

- 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: $\mu_x = 0$

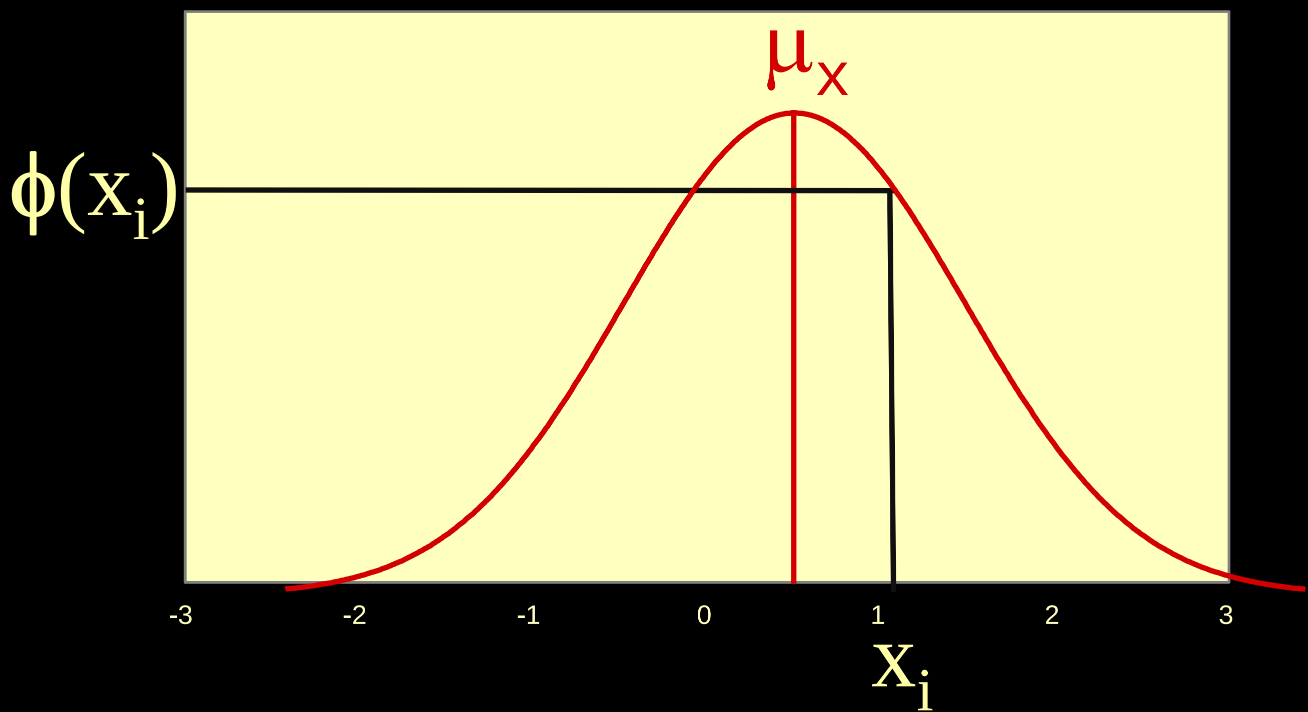
Probability density function



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

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

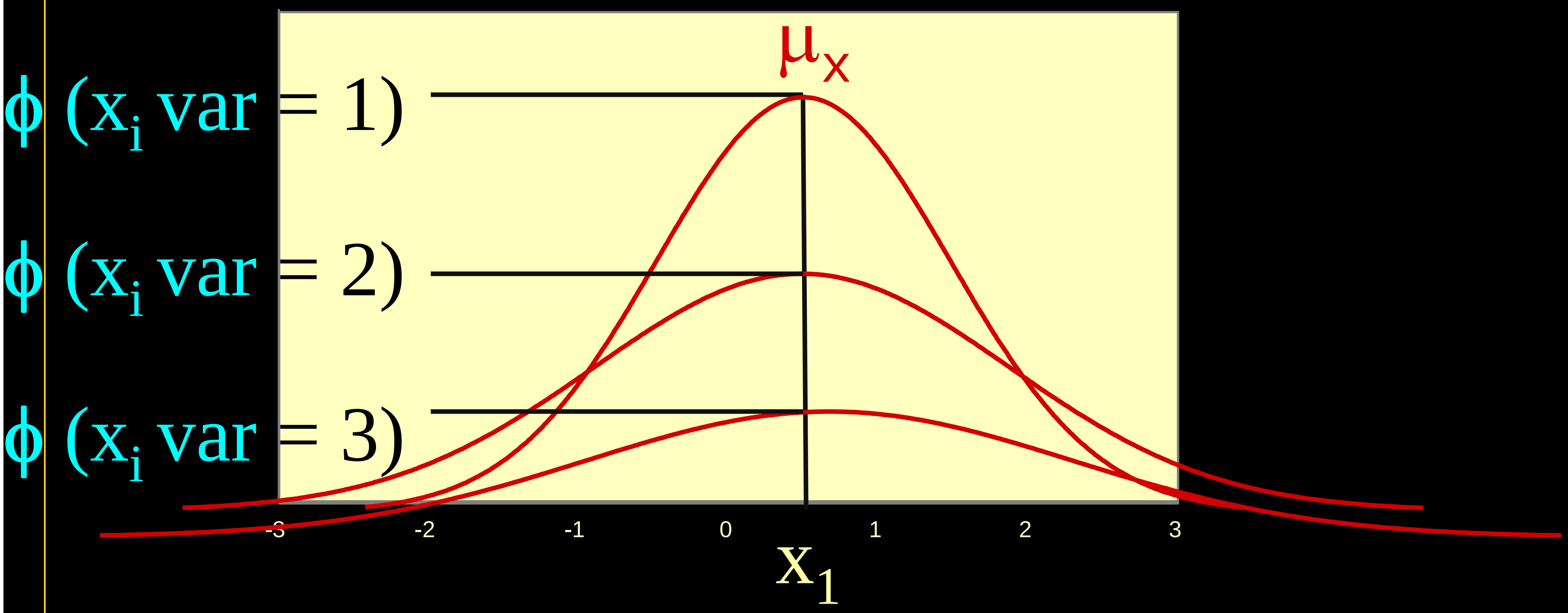
Function of *mean*



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

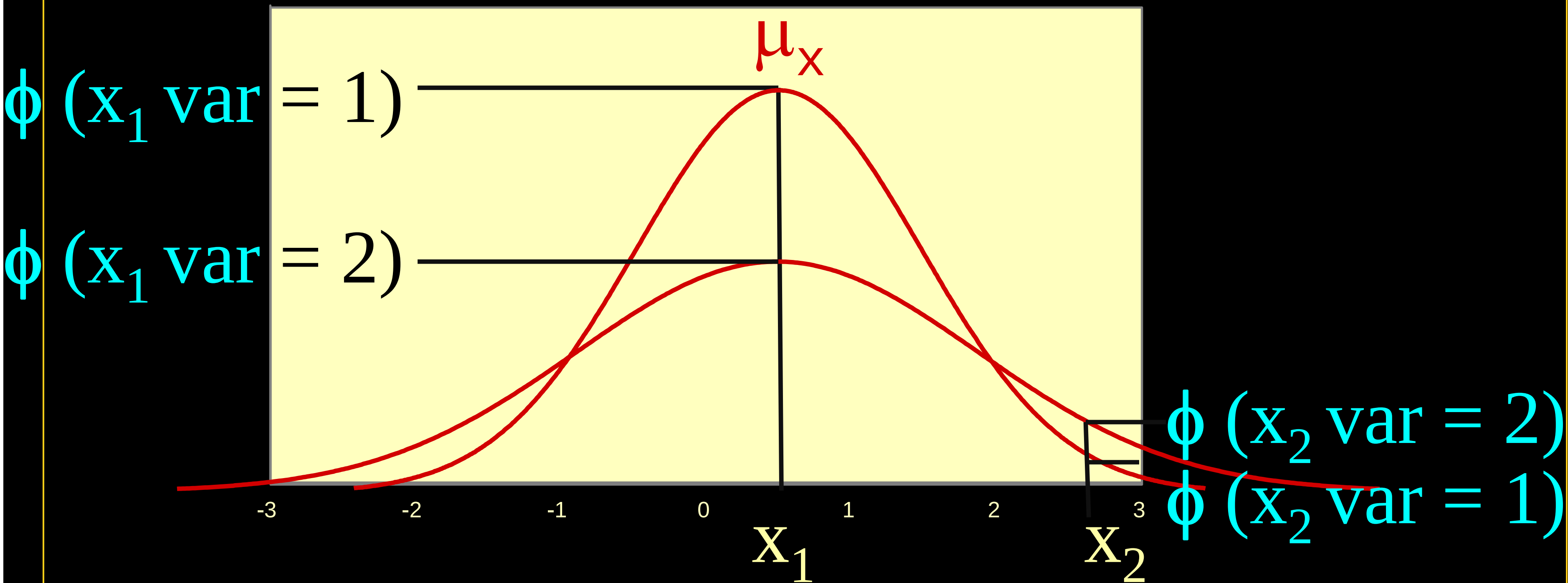
Height of normal curve at x_1

Function of *variance*



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

Height of normal curve at x_1 and x_2

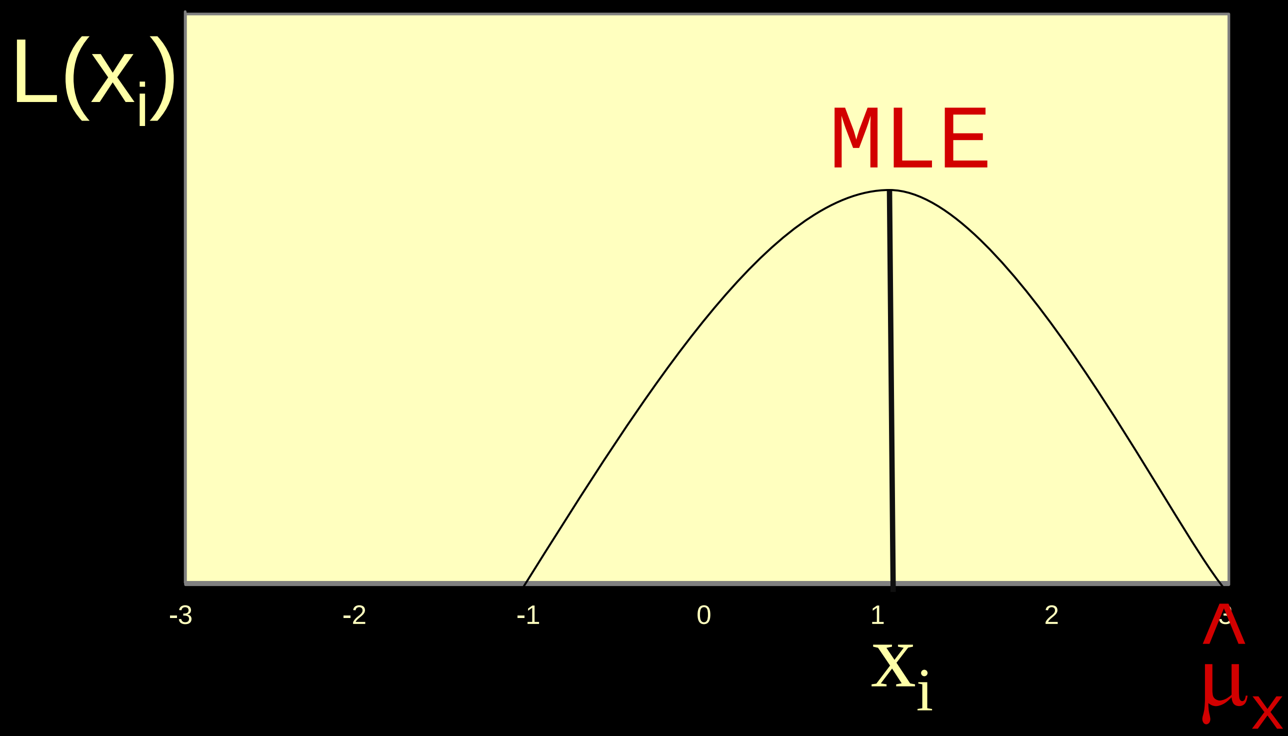


x_1 has higher likelihood with var=1 whereas x_2 has higher likelihood with var=2

VCU

Likelihood of x_i as a function of μ

Likelihood function



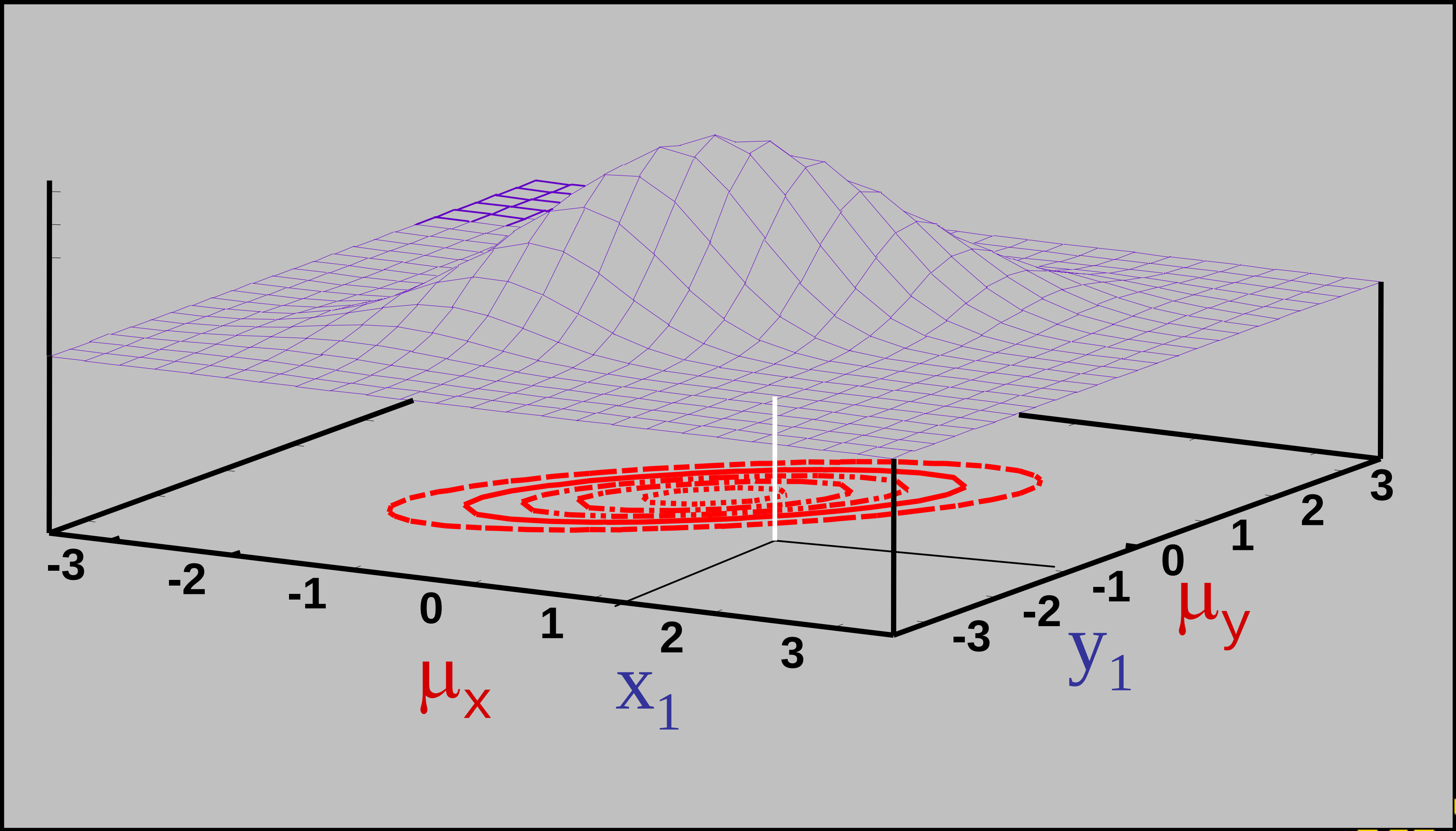
$L(x_i)$ is the likelihood of data point x_i for particular mean & variance estimates

Likelihood as a measure of “outlierness”

- Unlikely observation may be an outlier
  - Genuine
  - Data entry error
  - Model-specific
- Can use Mx feature to obtain case-wise likelihoods
  - Raw data
  - Option mx%p= uni_pi.out
  - Output for each case: the contribution to the -2ll as well as z-score statistic and Mahalanobis distance, weight and weighted likelihood
  - Generates R syntax to read in file, and sort by z-score
  - Beeby Medland & Martin (2006) ViewPoint and ViewDist: utilities for rapid graphing of linkage distributions and identification of outliers. *Behav Genet*. 2006 Jan;36(1):7-11

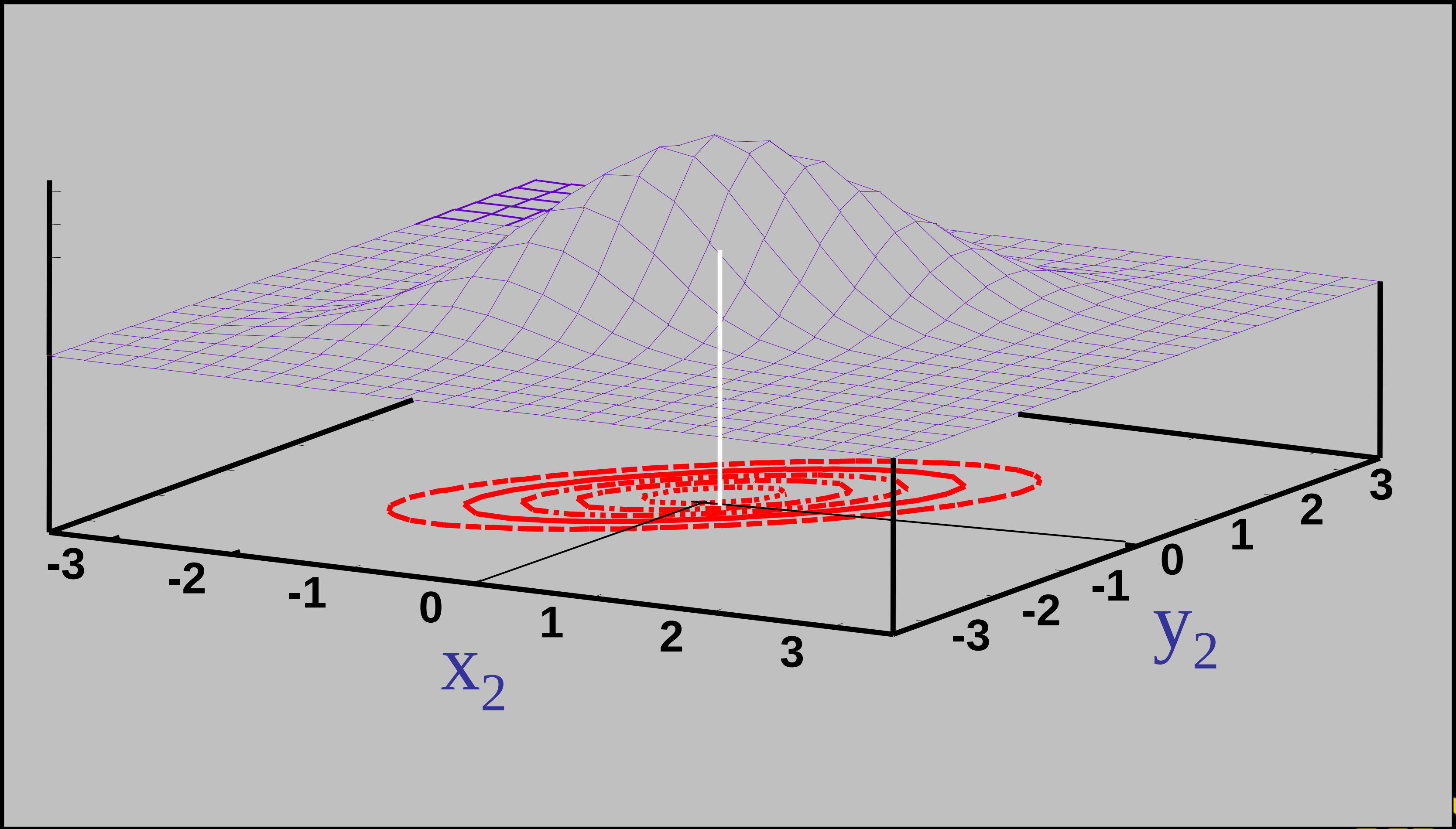
Height of bivariate normal density function

An unlikely pair of (x,y) values



Height of bivariate normal density function

A more likely pair of (x,y) values



Likelihood of Independent Observations

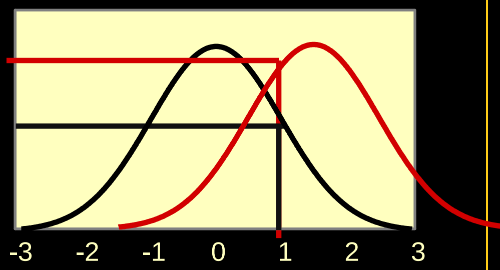
- Chance of getting two heads
- $L(x_1 \ldots x_n)$ = Product $(L(x_1),\ L(x_2)\ ,\ \ldots\ L(x_n))$
- $L(x_i)$ typically < 1

- 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 $-2\ln(L)$

VCU

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 χ_1^2

VCU

Exercises: Compute Normal PDF

- Get used to Mx script language
- Use matrix algebra
- Taste of likelihood theory

Mx script part 1: Declare groups and matrices

```
#NGroups 1

Title figure out likelihood by hand
Calculation
Begin Matrices;
 E symm 2 2 ! Expected Covariance Matrix
 H full 1 1 ! One half
 T full 1 1 ! Two
 M full 2 1 ! Mean vector
 P full 1 1 ! Pi
 X full 2 1 ! Observed Data
End Matrices;
```

Mx script part 2: Put values in matrices

Matrix E
 1 .0 1
Matrix H .5
Matrix M 0 0
Matrix P 3.141592
Matrix T 2
Matrix X 1 2

Mx script part 3: Matrix Algebra

```
Begin Algebra;
 O=T*P*\sqrt\det(E); ! Fractional part, 2pi*sqrt(det(e))
 Q=(X-M)'&(E~); ! Mahalanobis Distance
 R=\exp(-H*Q); ! e to the power -.5*Mahalanobis distance
 S=-T*\ln(R%O);! minus twice log-likelihood

 Z=-T*\ln(\pdfnor(X'_M'_E));  ! A simpler way
End Algebra;

End Group;
```

Exercises 1

- Bivariate normal distribution
  - Means [110.28   112.00]
  - Covariance matrix [299.40
    174.20   281.18]
- Compute likelihood of observed vector x = [87 89]

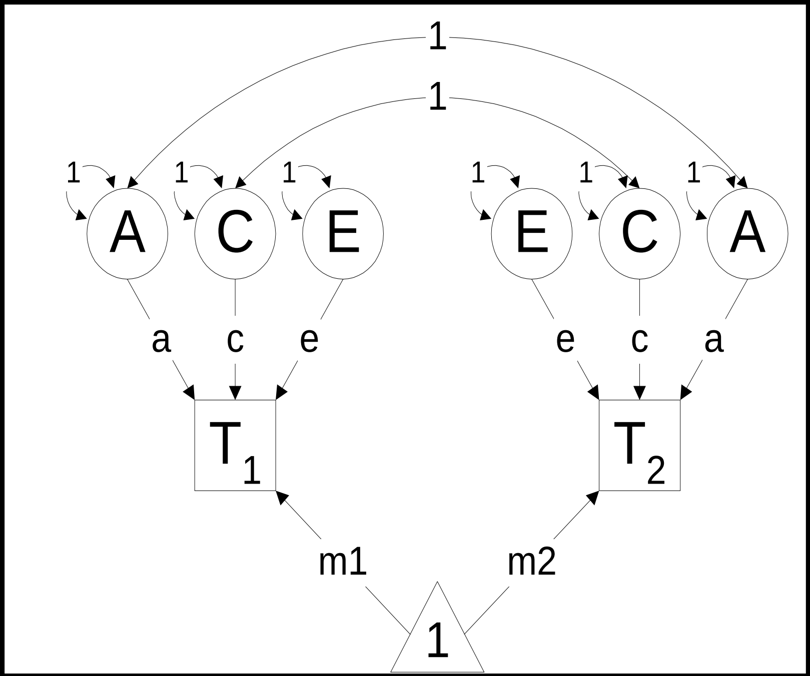
Exercises 2

- Bivariate normal distribution
  - Means [1 1]
  - Covariance matrix [1 .3
    .3 1]
- Compute likelihood of observed vector x = [1 2]
- Compute likelihood with correlation of .0 instead
- *Optional compute likelihood of observed vector x = [-2 -2] with correlations .5, .0, and 0*
- *Which is the most likely combination of model and data?*

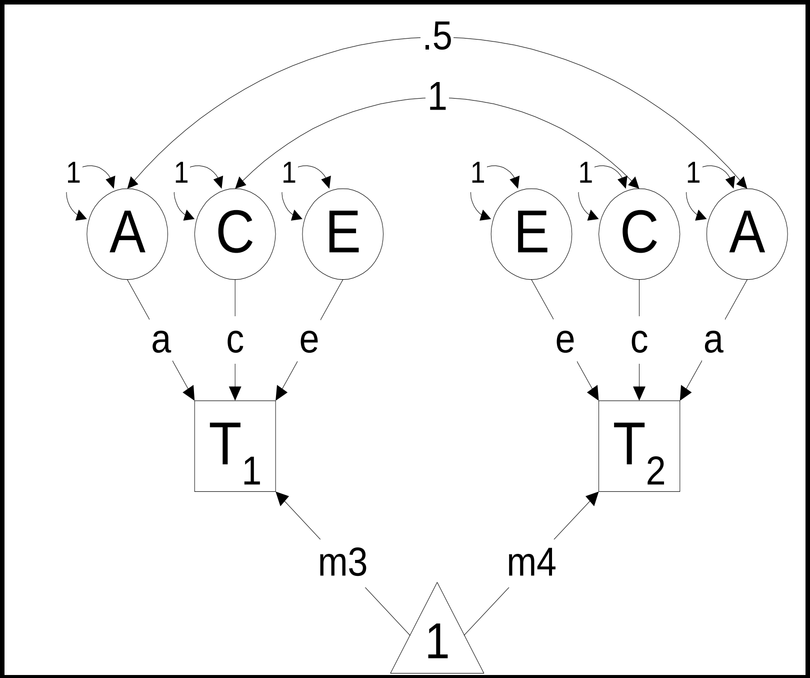
Exercises 3

- Univariate normal distribution
  - Mean [1]
  - Variance [1]
- Compute likelihood of observed vector x = [1]
- Compute likelihood of observed vector x = [2]
- Compute their product
- Which bivariate case does this equal?

Two Group Model: ACE



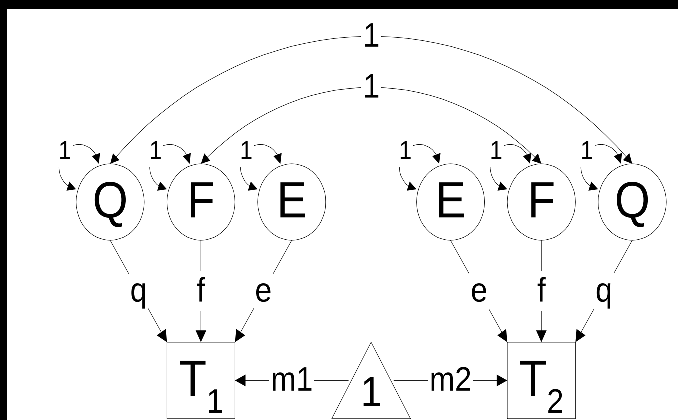
MZ twins

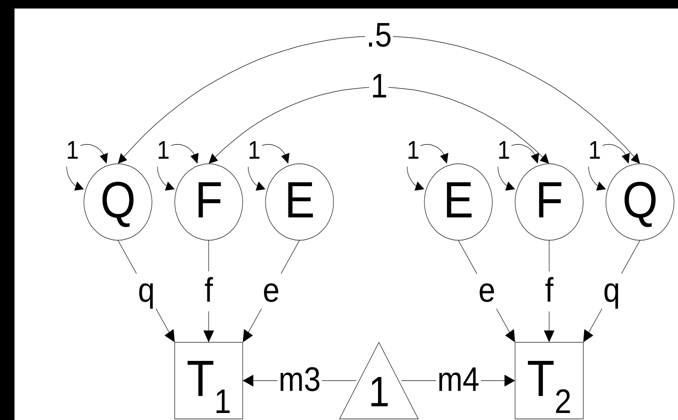


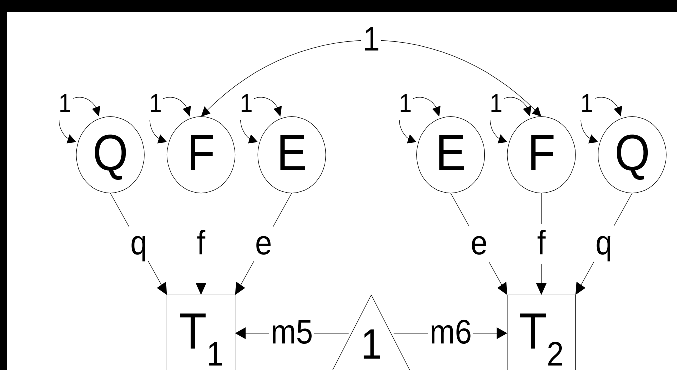
DZ twins

7 parameters

DZ by IBD status







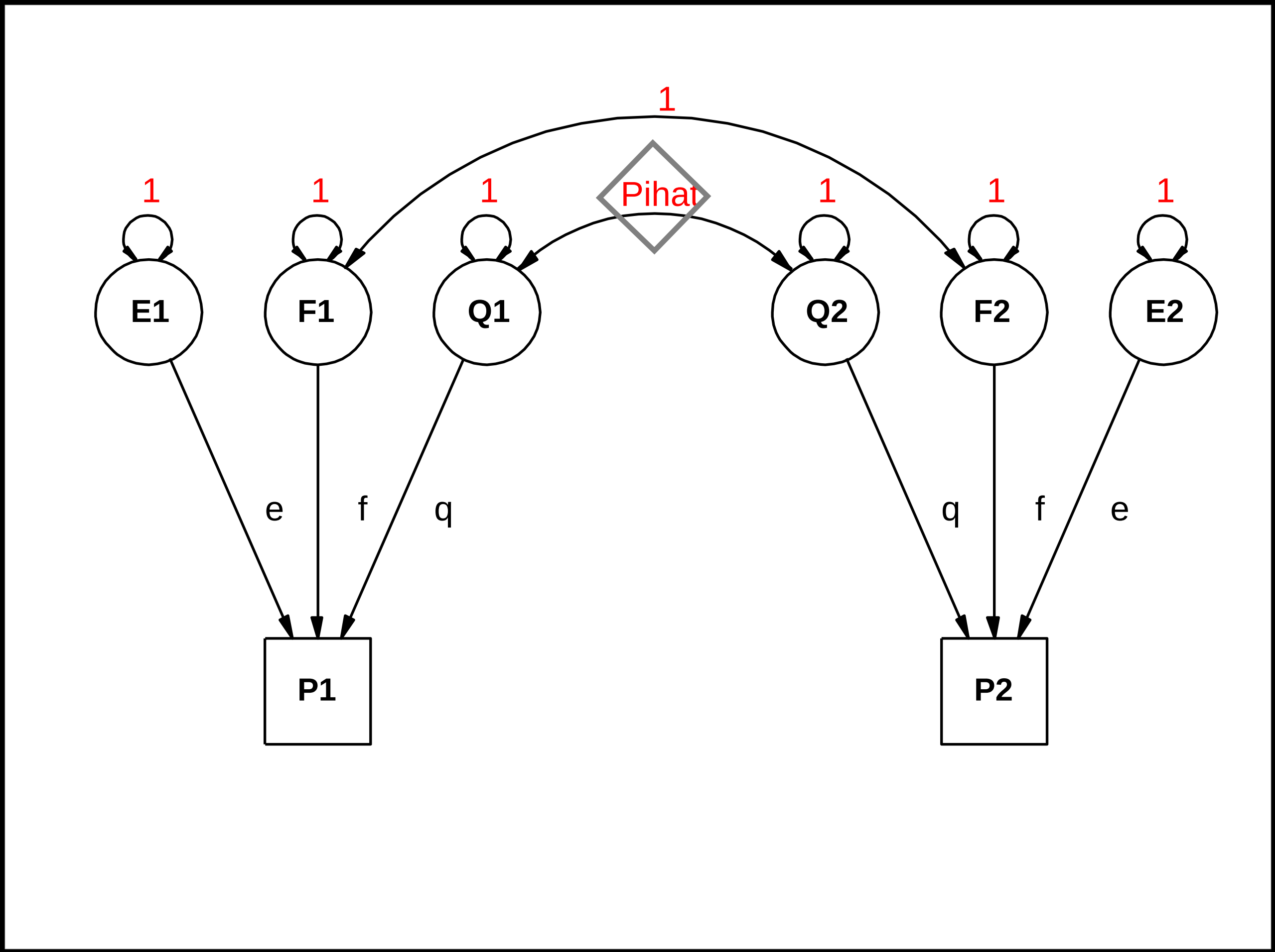
- Variance = Q + F + E
- Covariance = πQ + F + E

Extensions to More Complex Applications

- Endophenotypes
- Linkage Analysis
- Association Analysis

Basic Linkage (QTL) Model

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

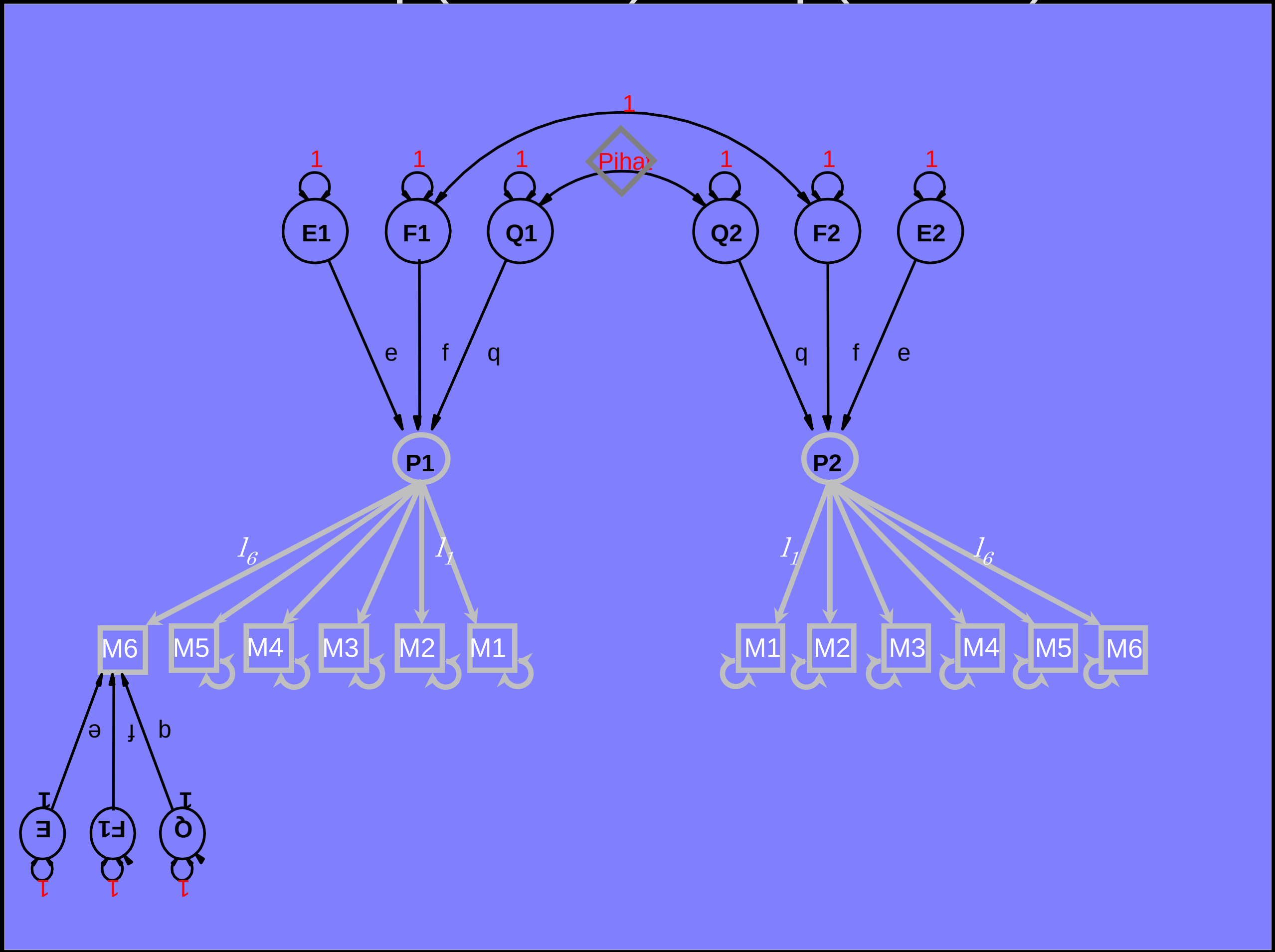


Q: QTL Additive Genetic F: Family Environment E: Random Environment

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

Measurement Linkage (QTL) Model

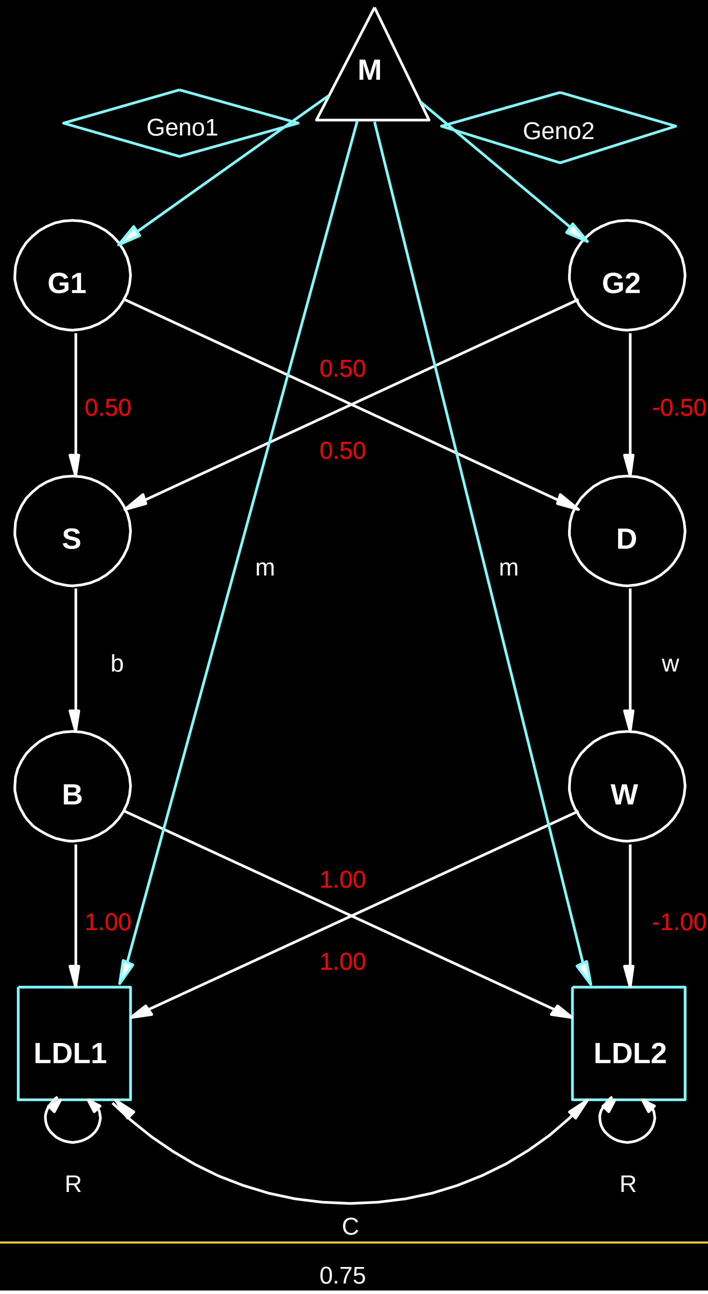
$\hat{\pi} = p(IBD=2) + .5\ p(IBD=1)$



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

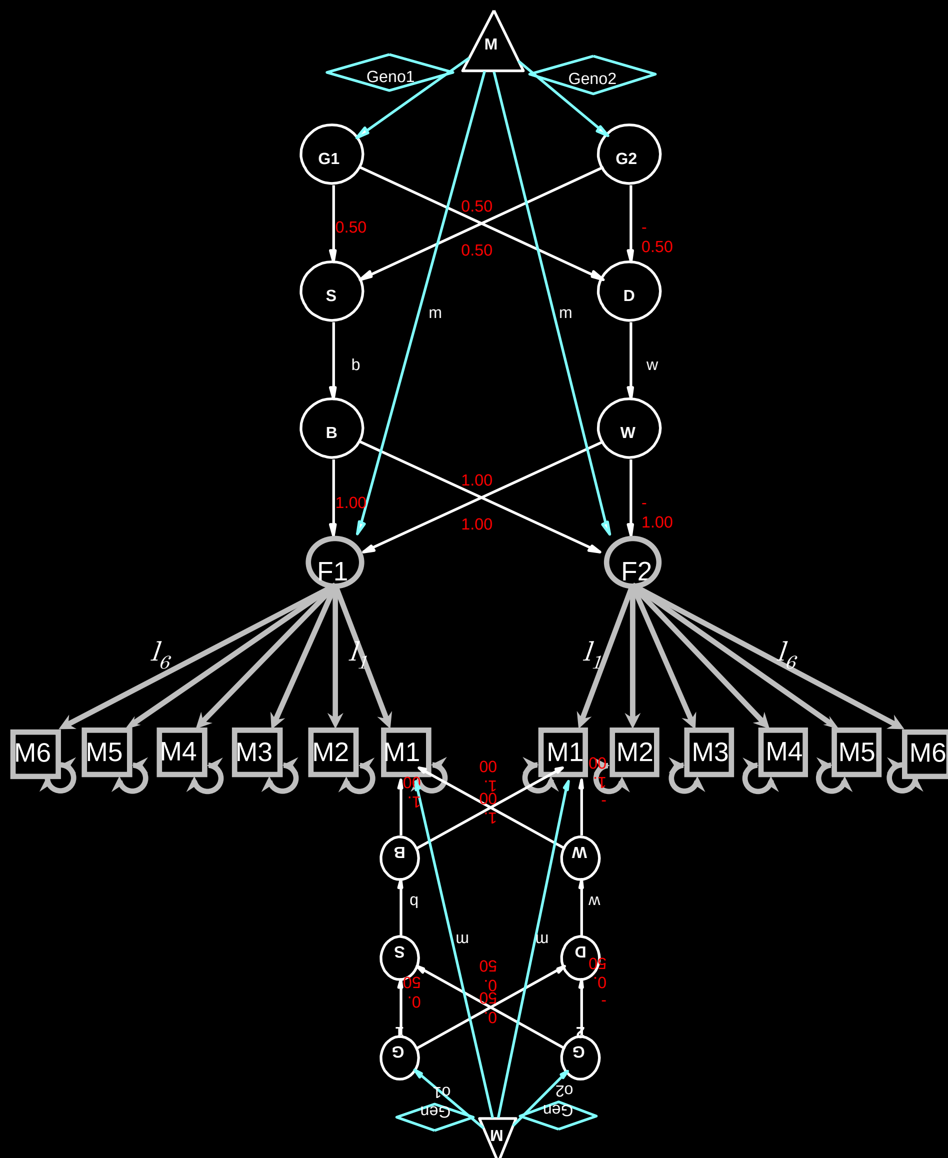
3 estimated parameters: q, f and e Every sibship may have different model

Fulker Association Model



Multilevel model
for the means

Measurement Fulker Association Model (SM)



Multivariate Linkage & Association Analyses

- Computationally burdensome
- Distribution of test statistics questionable
- Permutation testing possible
  - Even heavier burden
  - Sarah Medland's rapid approach
- Potential to refine both assessment and genetic models
- Lots of long & wide datasets on the way
  - Dense repeated measures: EMA; fMRI(!)
  - Need to improve software! Open source Mx

VCU