Ordinal Data

Brad Verhulst

&

Sarah Medland

Special thanks to Frühling Rijsdijk and those who came before

Analysis of ordinal variables

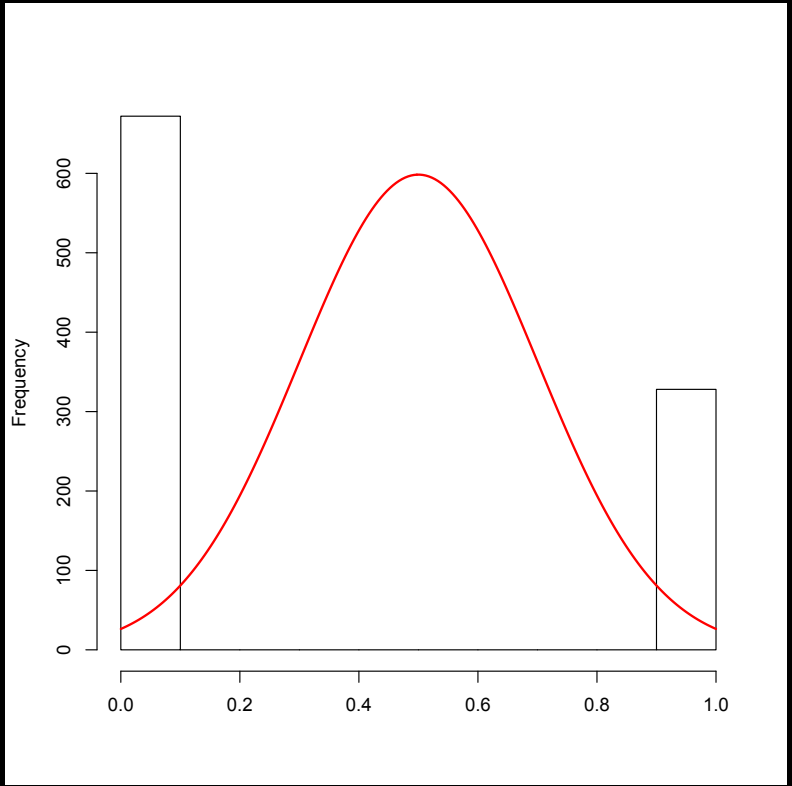
- The session aims to provide an intuition for how we estimate correlations from ordinal data (as twin analyses rely on covariance/correlation)
- For this we need to introduce the concept of ‘Liability’ or ‘liability threshold models’
- This is followed by a more mathematical description of the model

Ordinal data

- **Measuring instrument discriminates between two or a few ordered categories e.g.:**
  - Absence (0) or presence (1) of a disorder
  - Score on a single likert item
  - Number of symptoms
- In such cases the data take the form of counts, i.e. the number of individuals within each category of response

Problems with the treating binary variables as continuous

- Normality – Binary Variables are not distributed normally, ***obviously***.



- This means that the error terms cannot be normally distributed

Two Ways of Thinking about Binary Dependent Variables

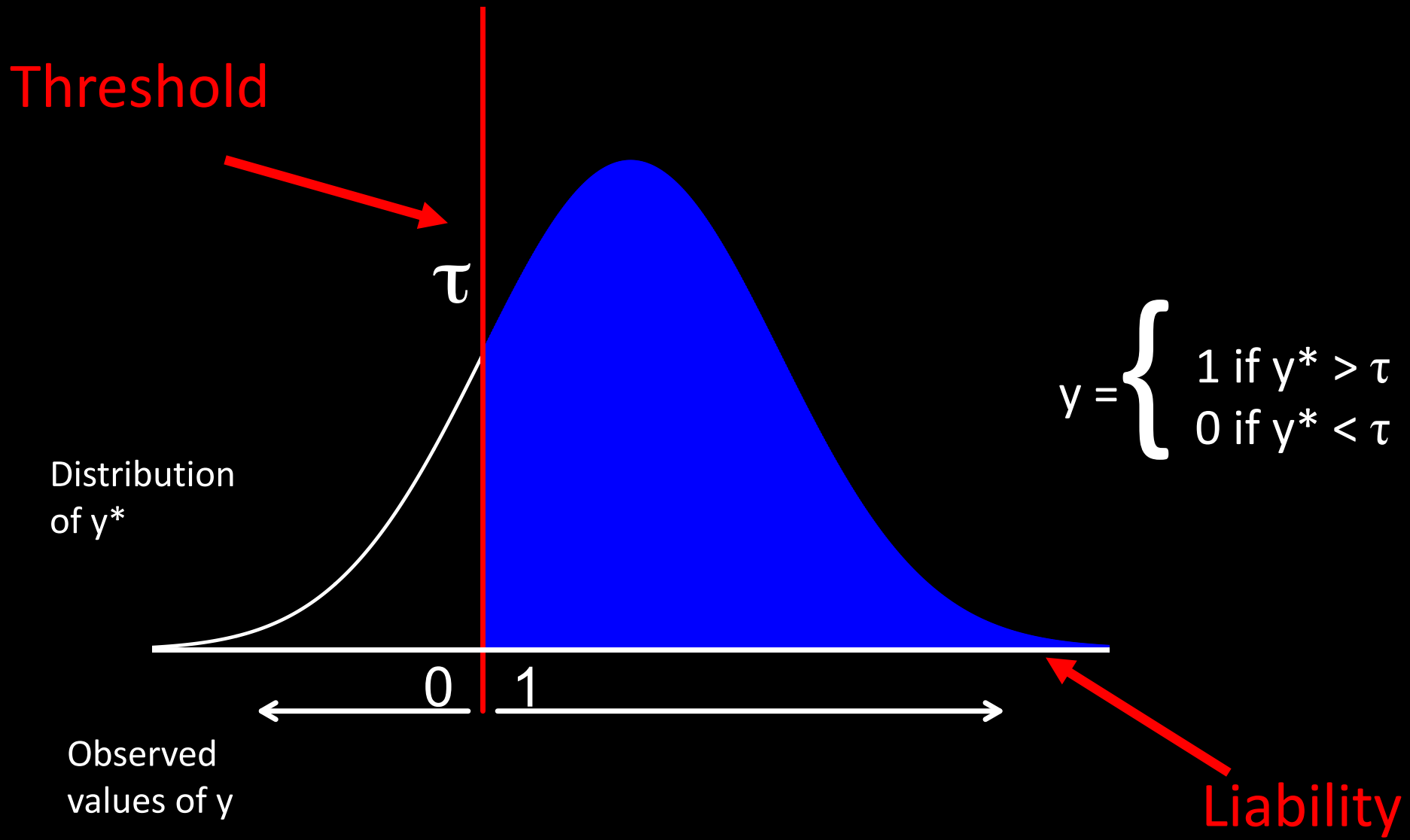
1. Assume that the observed binary variable is indicative of an underlying, latent (unobserved) continuous, normally distributed variable.
   - We call the unobserved variable a **Liability**

2. Assume the Binary Variable as a random draw from a Binomial (or Bernuilli) Distribution (Non-Linear Probability Model).

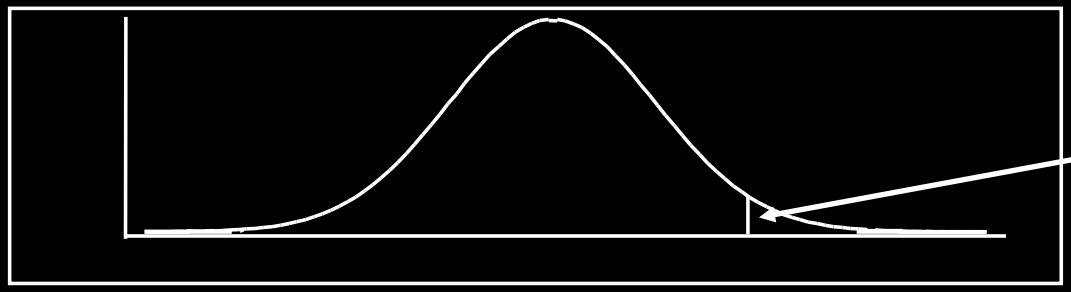
Binary Variables as indicators of Latent Continuous Variables

- Assume that the observed binary variable is indicative of an underlying, latent (unobserved) continuous, normally distributed variable.
    - Assumptions:
        1. Categories reflect an imprecise measurement of an underlying *normal distribution* of liability
        2. The liability distribution has 1 or more thresholds

Intuition behind the Liability Threshold Model (LTM)

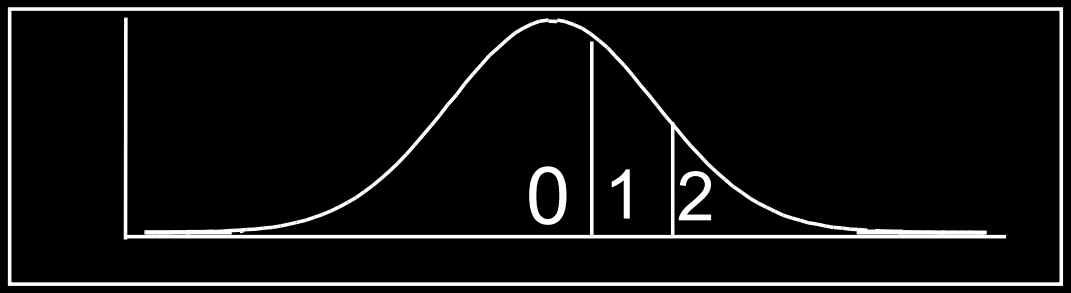


For disorders:



The ***risk*** or **liability** to a disorder is normally distributed
When an individual exceeds a threshold they have the disorder.
Prevalence: proportion of affected individuals.

For a single questionnaire item score e.g:



0 = not at all
1 = sometimes
2 = always

Does not make sense to talk about prevalence: we simply count the endorsements of each response category

Intuition behind the Liability Threshold Model (LTM)

- We can only observe binary outcomes, affected or unaffected, but people can be more or less affected.
- Since the variables are latent (and therefore not directly observed) we cannot estimate the means and variances we did for continuous variables.
- Thus, we have to make assumptions about them (pretend that they are some arbitrary value).

Identifying Assumptions

The traditional assumption

Mean Assumption

The intercept (mean) is 0

or

The threshold is 0 ($\tau = 0$)

- Either of these two assumptions provide equivalent model fit and the intercept is a transformation of τ.

Variance Assumption

The Probit Model

$Var(\varepsilon|x) = 1$ in the normal-ogive model

The Logit Model

$Var(\varepsilon|x) = \pi^2/3$ in the logit model.

Assumption 3

The conditional mean of ε is 0.

- This is the same assumption as we make for continuous variables, and allows the parameters to be unbiased

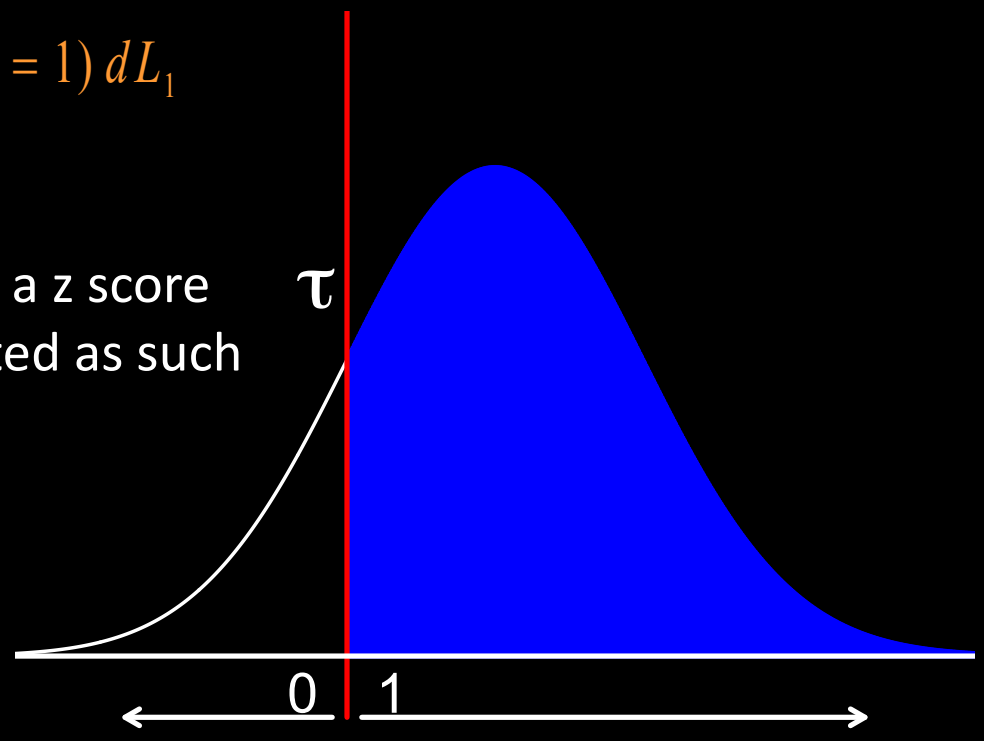
Identifying Assumptions of Ordinal Associations

- The assumptions are arbitrary.
  - We can make slightly different assumptions and still estimate the model, but that could come at the cost of efficiency, bias and ease of interpretation.
- The assumptions are necessary.
  - The magnitude of the covariance depends on the scale of the dependent variable.
  - If we don't make some assumptions about the variances, then the correlation coefficients are unidentified.

Intuitive explanation of thresholds in the univariate normal distribution

$$\int_{z_T}^{\infty} \Phi\left(L_1; \mu = 0, \sigma^2 = 1\right) dL_1$$

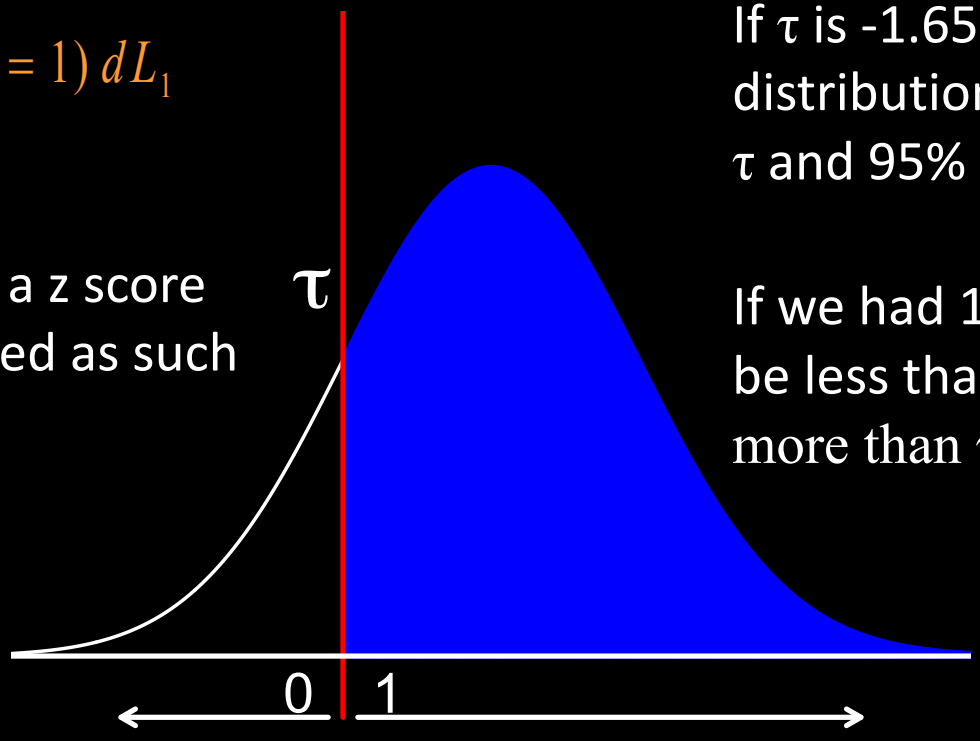
The threshold is just a z score and can be interpreted as such



Intuitive explanation of thresholds in the univariate normal distribution

$$\int_{z_T}^{\infty} \Phi\left(L_1; \mu = 0, \sigma^2 = 1\right) dL_1$$

The threshold is just a z score and can be interpreted as such



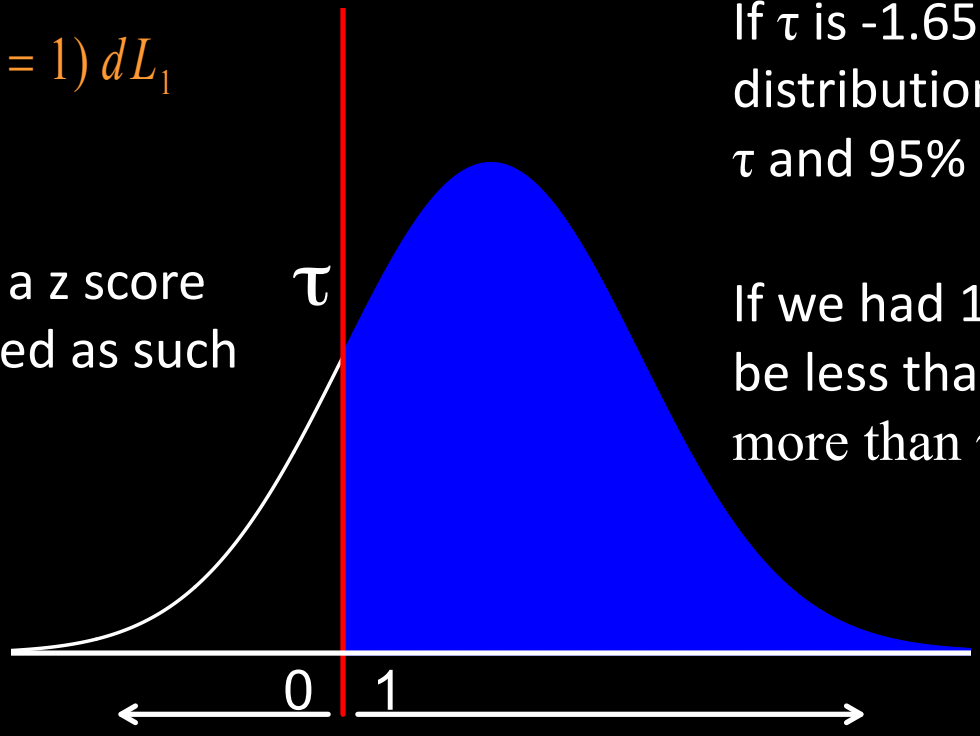
If τ is -1.65 then 5% of the distribution will be to the left of τ and 95% will be to the right

If we had 1000 people, 50 would be less than τ and 950 would be more than τ

Intuitive explanation of thresholds in the univariate normal distribution

$$\int_{z_T}^{\infty} \Phi\left(L_1; \mu = 0, \sigma^2 = 1\right) dL_1$$

The threshold is just a z score and can be interpreted as such



If τ is -1.65 then 5% of the distribution will be to the left of τ and 95% will be to the right

If we had 1000 people, 50 would be less than τ and 950 would be more than τ

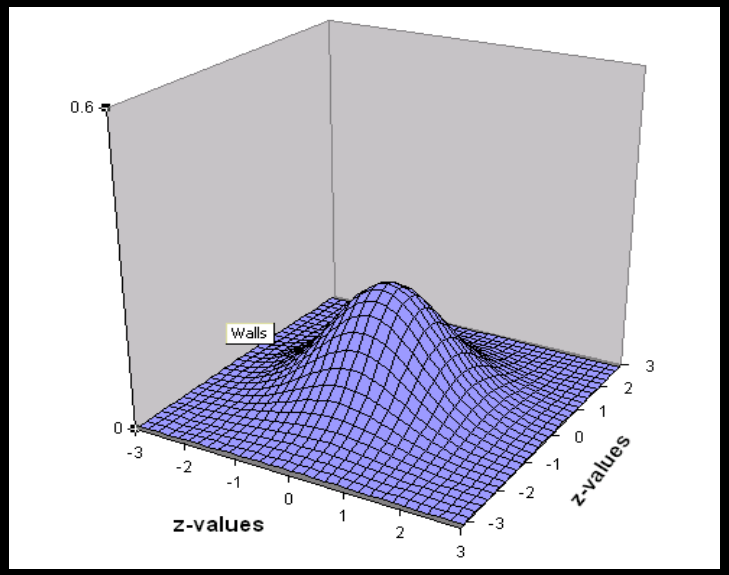
If τ is 1.96 then 97.5% of the distribution will be to the left of τ and .025% will be to the right

If we had 1000 people, 975 would be less than τ and 25 would be more than τ

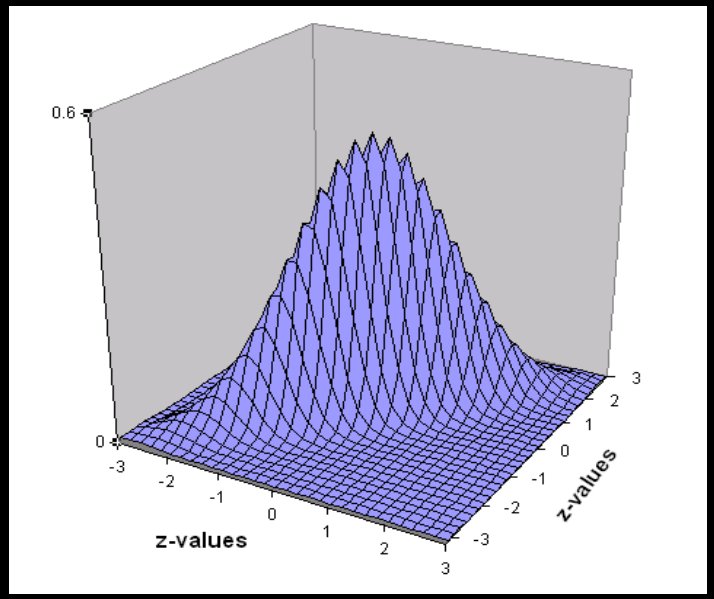
Bivariate normal distribution

- There are 2 variables
- We need to say something about the covariance

***r* =.00**



***r* =.90**



Two binary traits (e.g. data from twins)

Contingency Table with 4 observed cells:

Cell a: pairs concordant for unaffected
Cell b/c: pairs discordant for the disorder
Cell d: pairs concordant for affected

Twin1 / Twin2	0	1
0	a	b
1	c	d

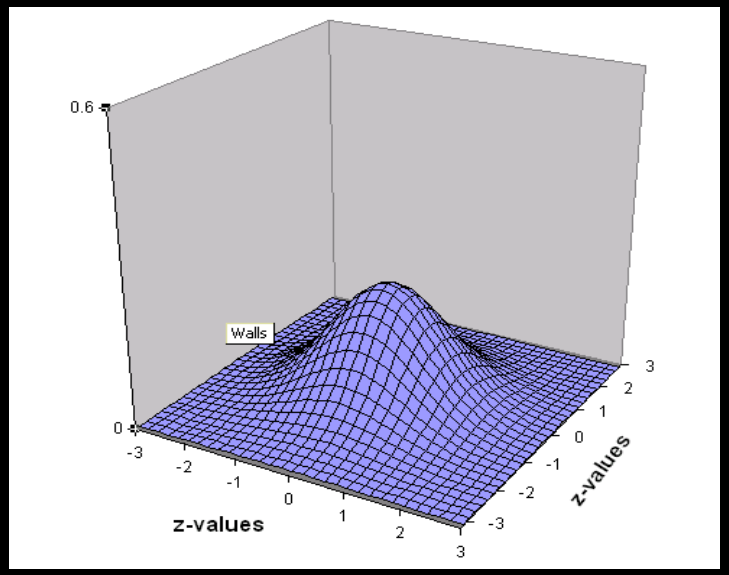
0 = unaffected
1 = affected

Joint Liability Threshold Model for twin pairs

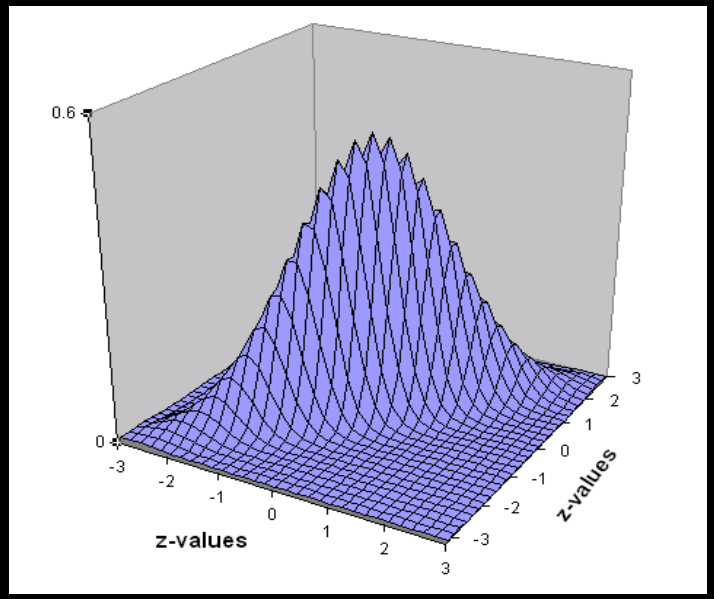
- Assumed to follow a **bivariate normal** distribution, where both traits have a mean of 0 and standard deviation of 1, and the **correlation** between them is what we want to know.

- The **shape** of a bivariate normal distribution is determined by the **correlation** between the traits

***r* =.00**

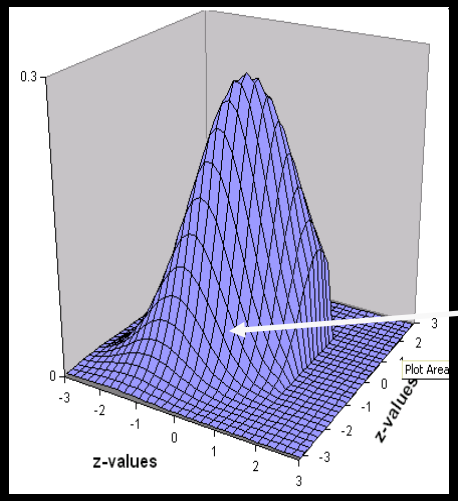


***r* =.90**

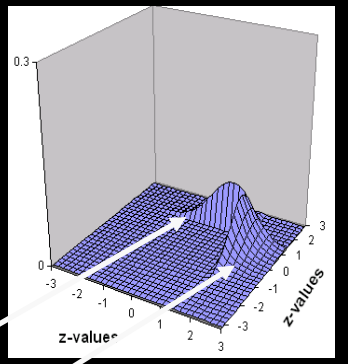


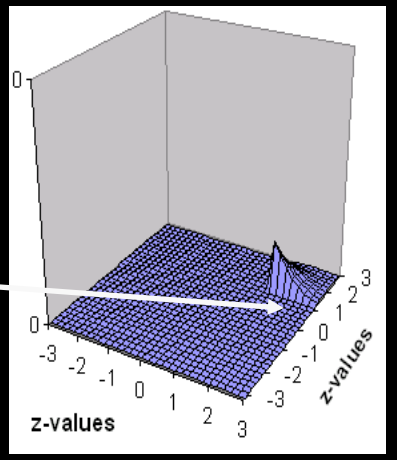
- The observed cell proportions relate to the proportions of the Bivariate Normal Distribution with a certain correlation between the latent variables (y_1 and y_2), each cut at a certain threshold

i.e. the joint probability of a certain response combination is the volume under the BND surface bounded by appropriate thresholds on each liability



y1 \ *y2*	0	1
0	00	01
1	10	11

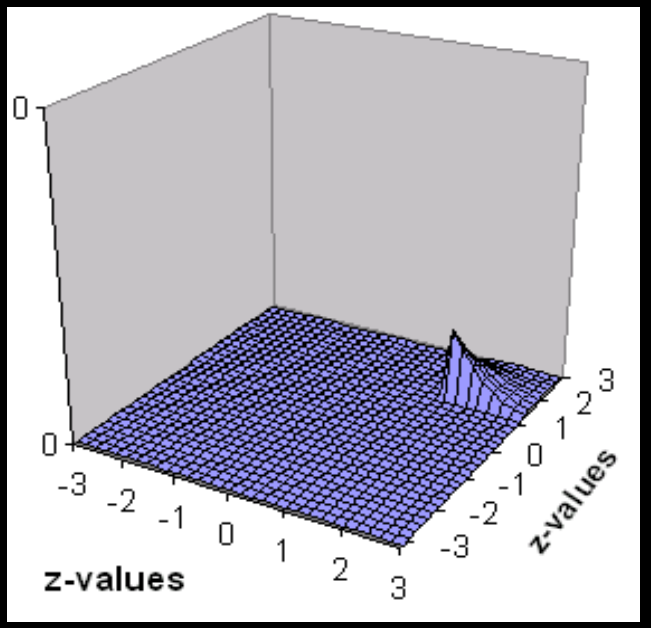




To calculate the cell proportions we rely on **Numerical integration** of the Bivariate Normal Distribution over the two liabilities

e.g. the probability that both twins are above T*c* :

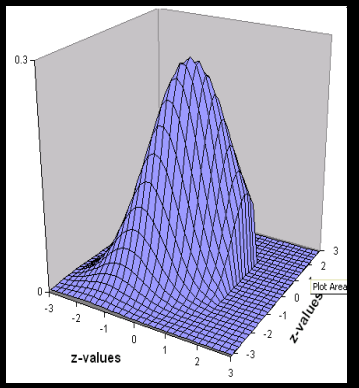
$$\int_{T_{c1}}^{\infty}\int_{T_{c2}}^{\infty}\Phi(y_1, y_2; \mu = 0, \Sigma)dy_1 dy_2$$



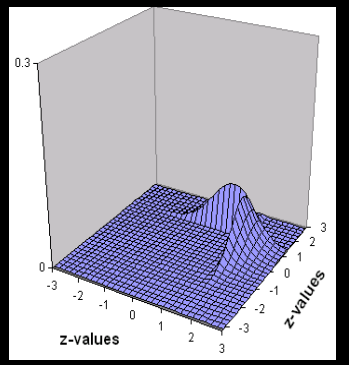
Φ is the bivariate normal probability density function,
y_1 and y_2 are the liabilities of twin1 and twin2,
with means of 0, and Σ the correlation between the two liabilities
T_{c1} is threshold (z-value) on y_1, T_{c2} is threshold (z-value) on y_2

Expected cell proportions

$$\int_{-\infty}^{T_{c1}} \int_{-\infty}^{T_{c2}} \Phi(y_1, y_2; \mu = 0, \Sigma) dy_1 dy_2$$



$$\int_{-\infty}^{T_{c1}} \int_{T_{c2}}^{\infty} \Phi(y_1, y_2; \mu = 0, \Sigma) dy_1 dy_2$$



$$\int_{T_{c1}}^{\infty} \int_{-\infty}^{T_{c2}} \Phi(y_1, y_2; \mu = 0, \Sigma) dy_1 dy_2$$

Estimation of Correlations and Thresholds

- Since the Bivariate Normal distribution is a known mathematical distribution, for each correlation (∑) and any set of thresholds on the liabilities we know what the expected proportions are in each cell.

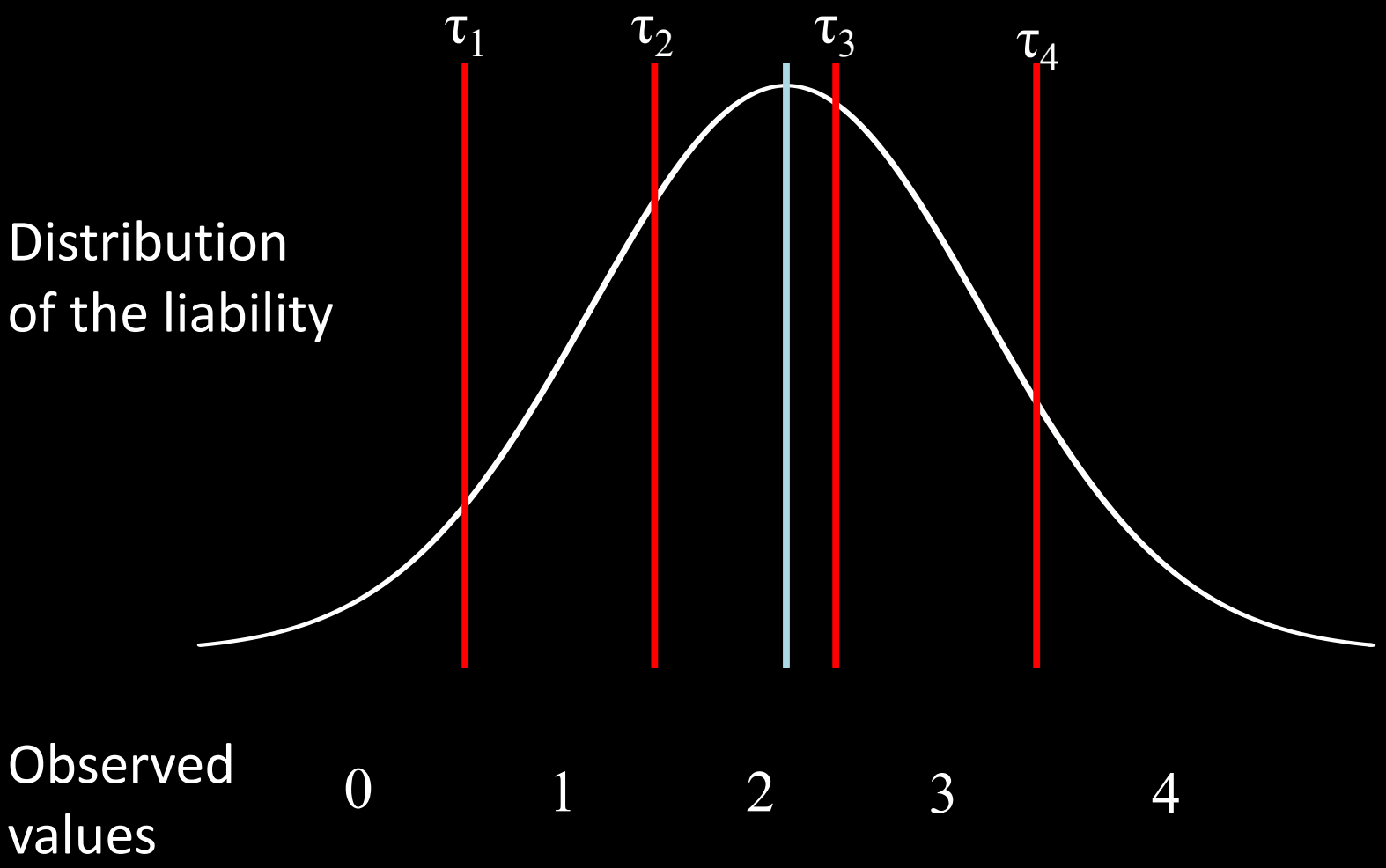
- Therefore, observed cell proportions of our data will inform on the most likely correlation and threshold on each liability.

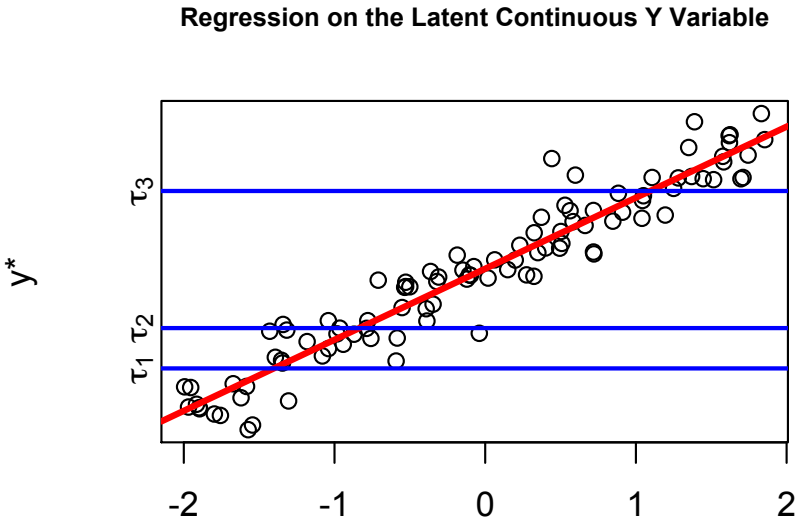
y1 \ y2	0	1
0	.87	.05
1	.05	.03

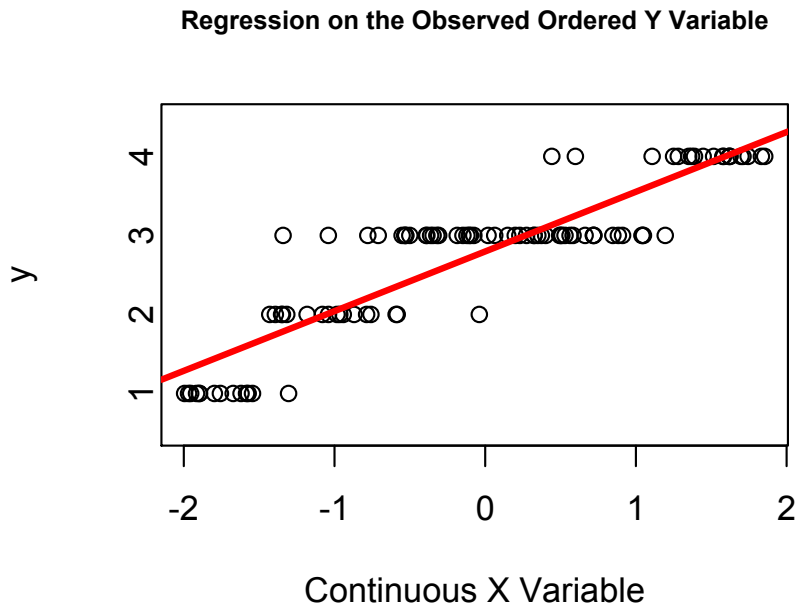
r = 0.60

$T_{c1}=T_{c2}$ = 1.4 (z-value)

Intuition behind the Liability Threshold Model with Multiple Cutpoints







Comparison between the regression of the latent y* and the observed y

It is important to keep in mind that the scale of the ordinal variable is arbitrary, and therefore it is virtually impossible to compare the slopes of the two graphs (even though they look pretty similar)

Psychological Methods
2004, Vol. 9, No. 3, 301–333

Copyright 2004 by the American Psychological Association
1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.3.301

Squeezing Interval Change From Ordinal Panel Data: Latent Growth Curves With Ordinal Outcomes

Paras D. Mehta
University of Illinois at Chicago

Michael C. Neale
Virginia Commonwealth University

Brian R. Flay
University of Illinois at Chicago

What happens if we change the default assumptions?

Mean Assumption

The intercept (mean) is 0

or

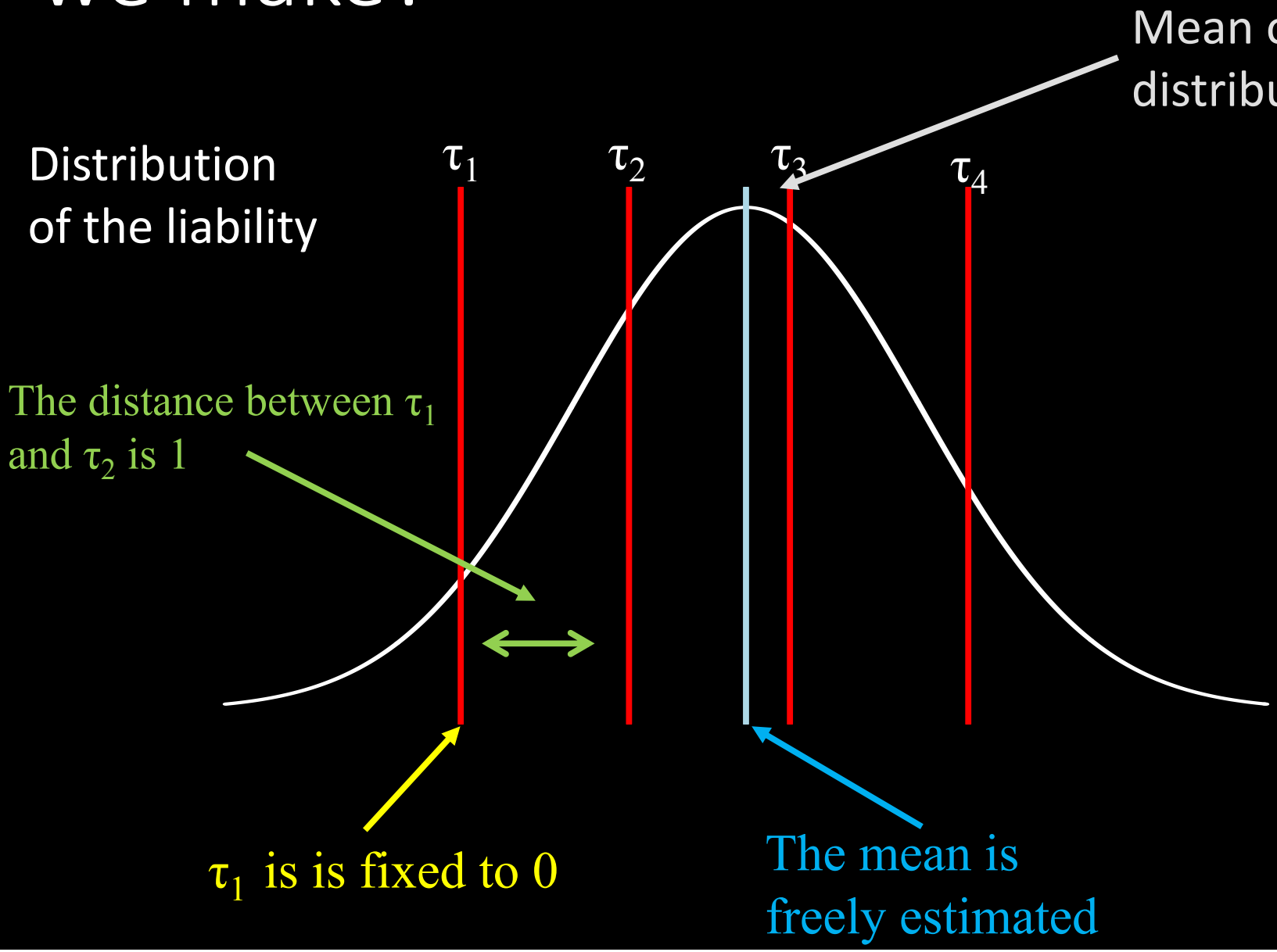
The threshold is 0 ($\tau = 0$)

Variance Assumption

$Var(\varepsilon|x) = 1$ in the normal-ogive model

Remember that we can make slightly different assumptions with equal model fit

What alternative assumptions could we make?



Psychological Methods
2004, Vol. 9, No. 3, 301–333

Copyright 2004 by the American Psychological Association
1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.3.301

Squeezing Interval Change From Ordinal Panel Data: Latent Growth Curves With Ordinal Outcomes

Paras D. Mehta
University of Illinois at Chicago

Michael C. Neale
Virginia Commonwealth University

Brian R. Flay
University of Illinois at Chicago

It is important to reiterate that the model fit is the same and that all the parameters can be transformed from one set of assumptions to another.

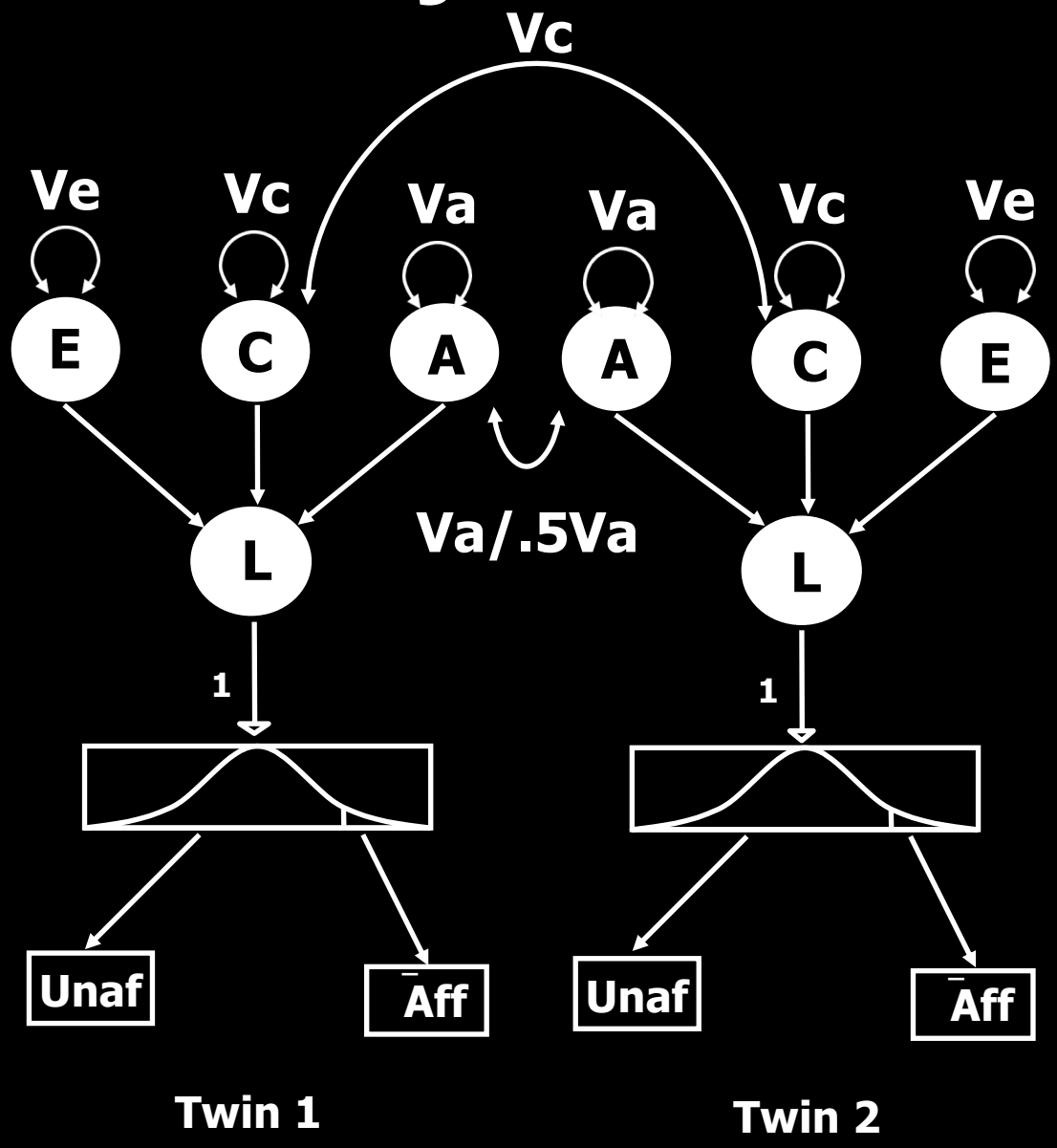
Bivariate Ordinal Likelihood

- The likelihood for each observed ordinal response pattern is computed by the expected proportion in the corresponding cell of the bivariate normal distribution
- The maximum-likelihood equation for the whole sample is the sum of -2* log of of the likelihood of each row of data (e.g. twin pairs)
- This -2LL is minimized to obtain the maximum likelihood estimates of the correlation and thresholds
- Tetra-choric correlation if *y*1 and *y*2 reflect 2 categories (1 Threshold); Poly-choric when >2 categories per liability

Twin Models

- Estimate correlation in liabilities separately for MZ and DZ pairs from contingency table
- Variance decomposition (A, C, E) can be applied to the *liability* of the trait
- Correlations in liability are determined by path model
- Estimate of the heritability of the ***liability***

ACE Liability Model



Variance constraint

Threshold model

Summary

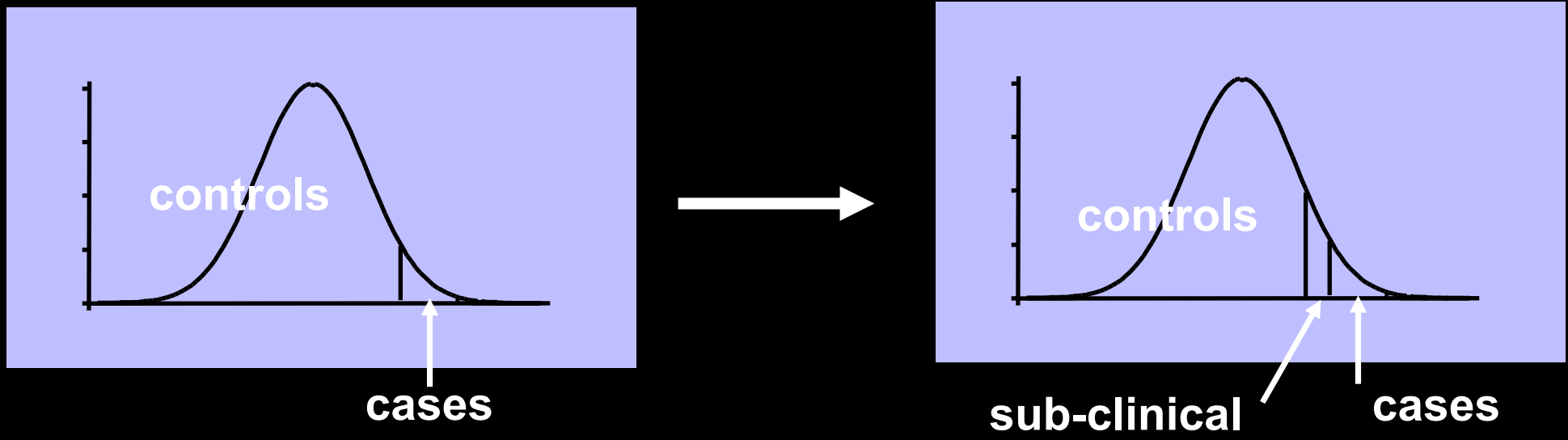
- OpenMx models ordinal data under a threshold model
- Assumptions about the (joint) distribution of the data (Standard Bivariate Normal)
- The relative proportions of observations in the cells of the Contingency Table are translated into proportions under the Multivariate Normal Distribution
- The most likely thresholds and correlations are estimated
- Genetic/Environmental variance components are estimated based on these correlations derived from MZ and DZ data

Power issues

- Ordinal data / Liability Threshold Model: less power than analyses on continuous data

Neale, Eaves & Kendler 1994

- Solutions:

1. Bigger samples
2. Use more categories

Please do not categorize continuous variables