ORDINAL DATA

BRAD VERHULST

TEXAS A&M UNIVERSITY

Special thanks to Frühling Rijsdijk,
Sarah Medland, Mike Neale and
all those who came before

ANALYSIS OF ORDINAL VARIABLES

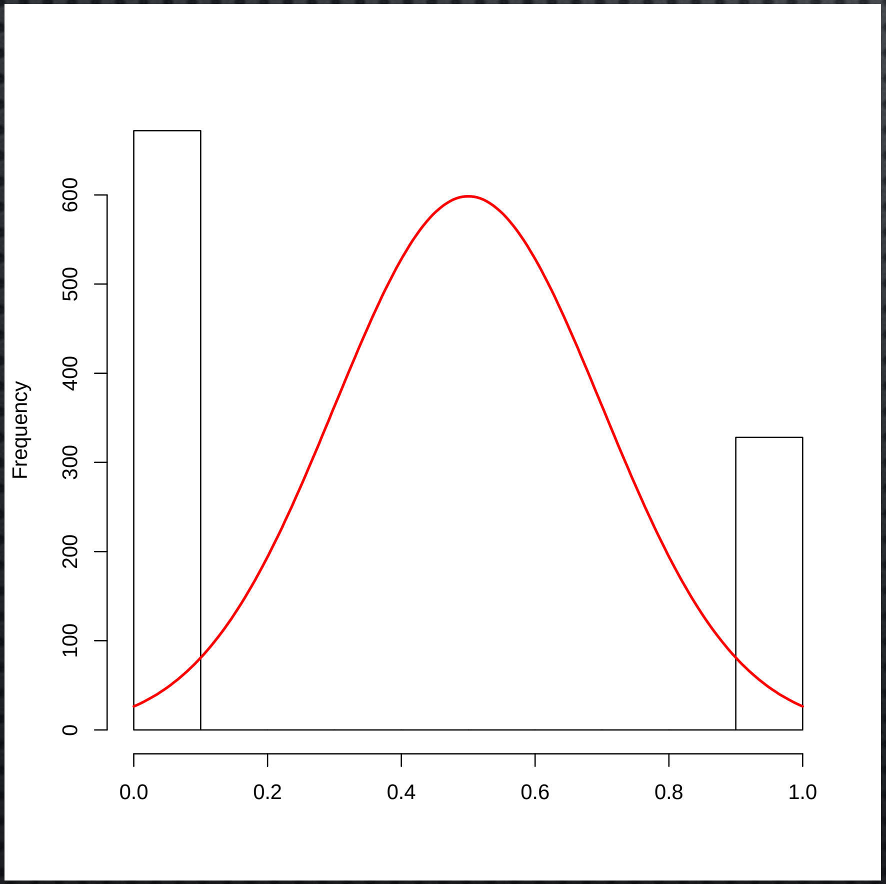
- Provide intuitive sense of how we estimate correlations from ordinal data
- Introduce the concept of liability threshold models
- Provide a mathematical description of the model

ORDINAL DATA

- We often measure behaviors using a limited number of <u>ordered</u> categories:
    - Absence (0) or presence (1) of a disorder
    - Severity of a disorder
    - Score on a single Likert item ‘none/some/lots’
    - Number of symptoms (far from ideal)

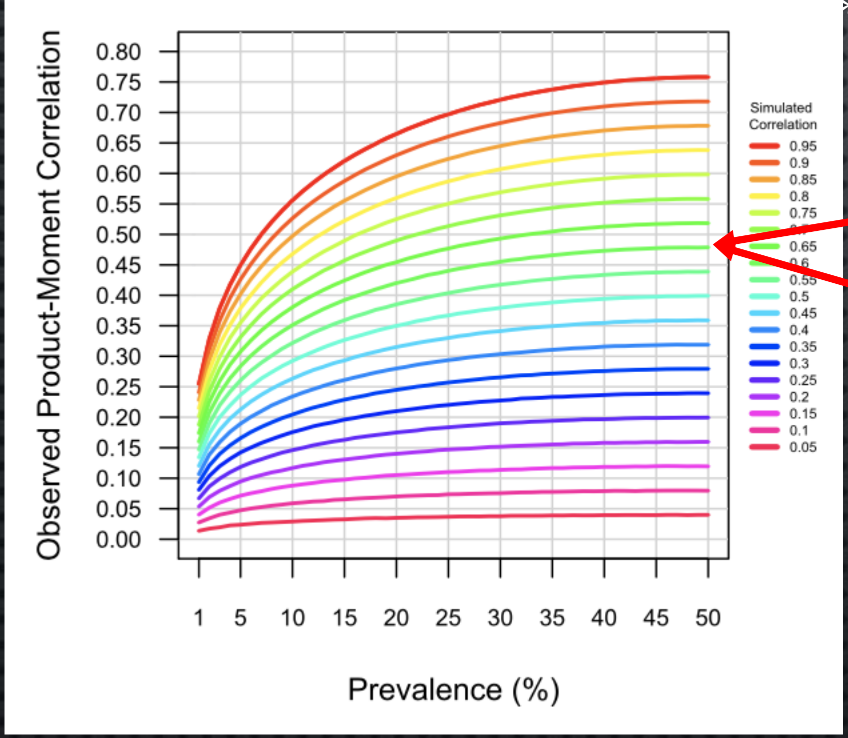
ORDINAL DATA REQUIRES DIFFERENT STATISTICAL METHODS

PROBLEMS WITH THE TREATING BINARY VARIABLES AS CONTINUOUS

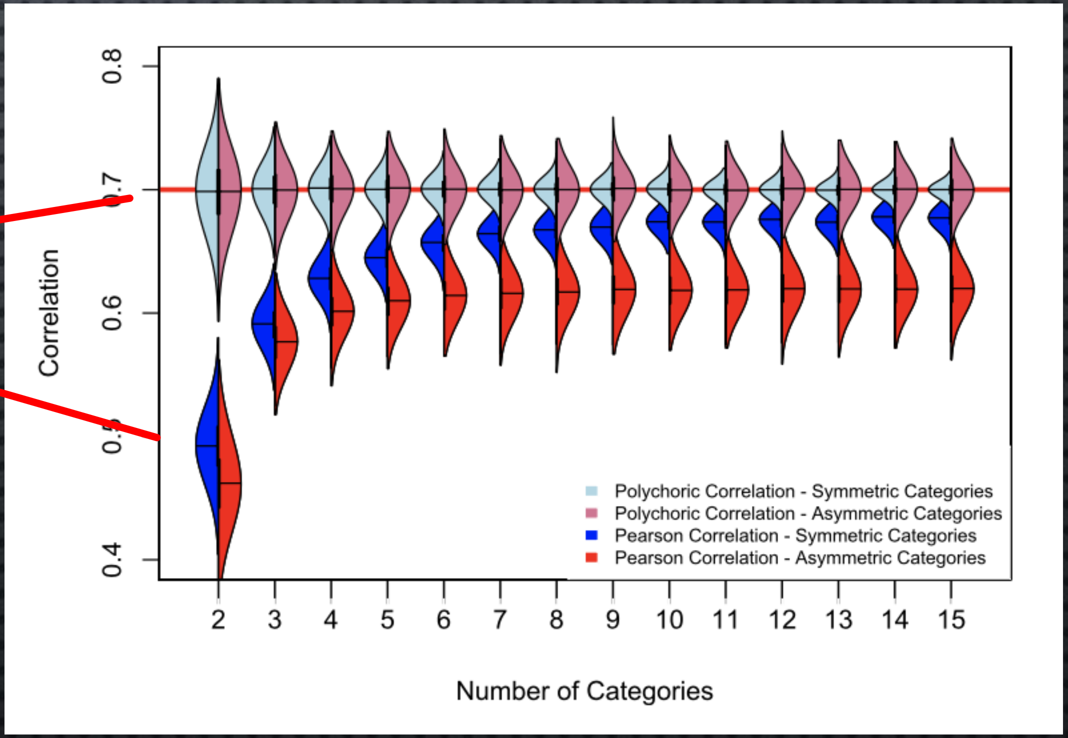


- NORMALITY – ORDINAL VARIABLES ARE NOT DISTRIBUTED NORMALLY, ***OBVIOUSLY***.
- THIS MEANS THAT THE ERROR TERMS CANNOT BE NORMALLY DISTRIBUTED

WHAT HAPPENS IF WE USE CONTINUOUS METHODS ON BINARY VARIABLES ?



*



Taken from Verhulst & Neale, 2021

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 (LATENT) VARIABLE A **LIABILITY**

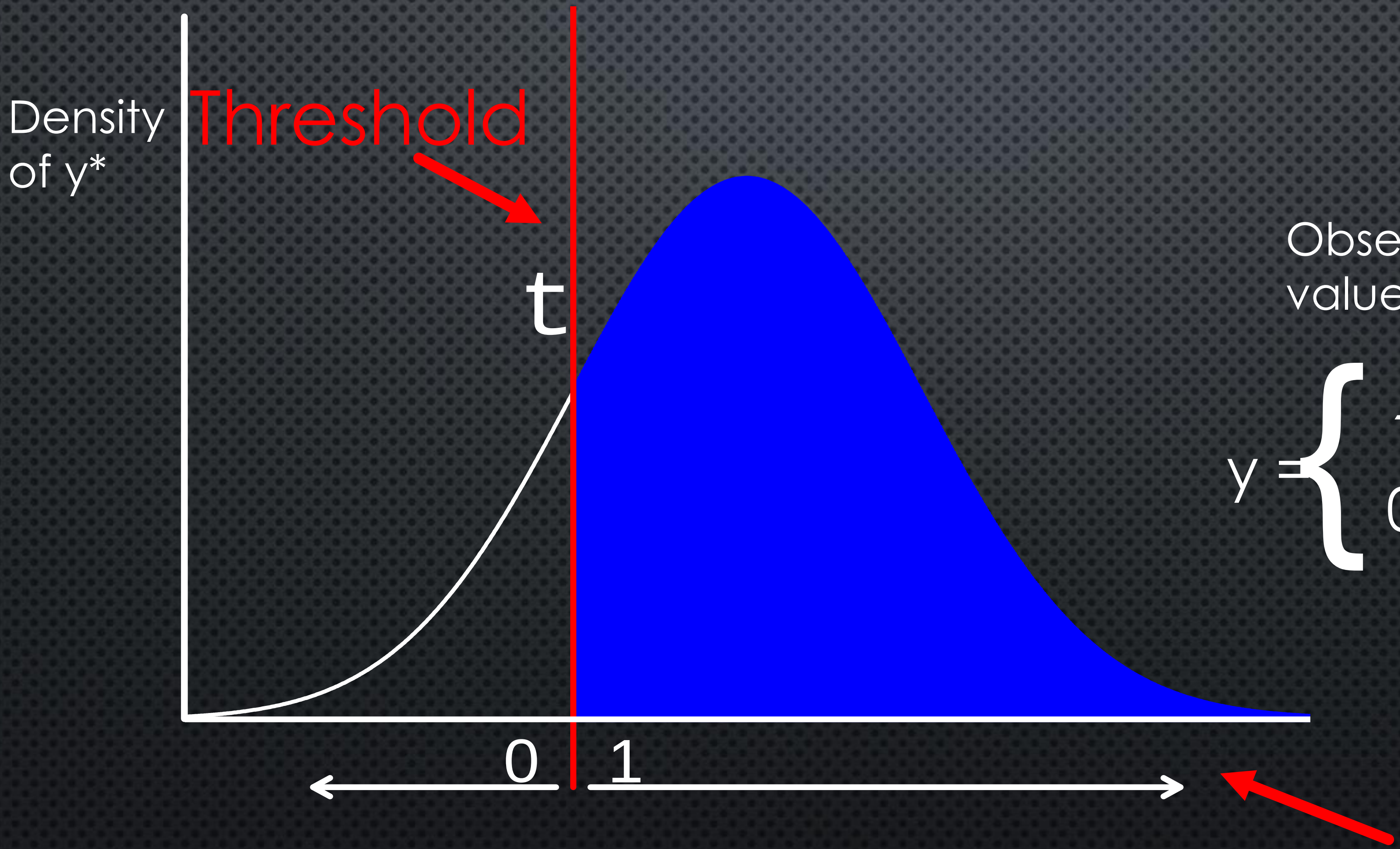
2. ASSUME THE BINARY VARIABLE AS A RANDOM DRAW FROM A BINOMIAL (OR BERNOUILLI) DISTRIBUTION (NON-LINEAR PROBABILITY MODEL). GENUINELY CATEGORICAL RESPONSES, NO UNDERLYING CONTINUOUS DISTRIBUTION.

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. This liability is thought to be influenced by many many things, each of which does almost nothing. The Central Limit Theorem predicts that variation should be distributed according to the normal or Gaussian distribution.
    2. The liability distribution has 1 or more thresholds

FUNDAMENTALS OF THE LIABILITY THRESHOLD MODEL



Observed values of y

$$y = \begin{cases} 1 \text{ if } y^* > \tau \\ 0 \text{ if } y^* < \tau \end{cases}$$

Liability Dimension

IDEAS BEHIND THE LIABILITY THRESHOLD MODEL (LTM)

- We only observe binary outcomes, affected or unaffected, but we know that 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

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

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

The Logit Model

Assumption 3

The conditional mean of e is 0.

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

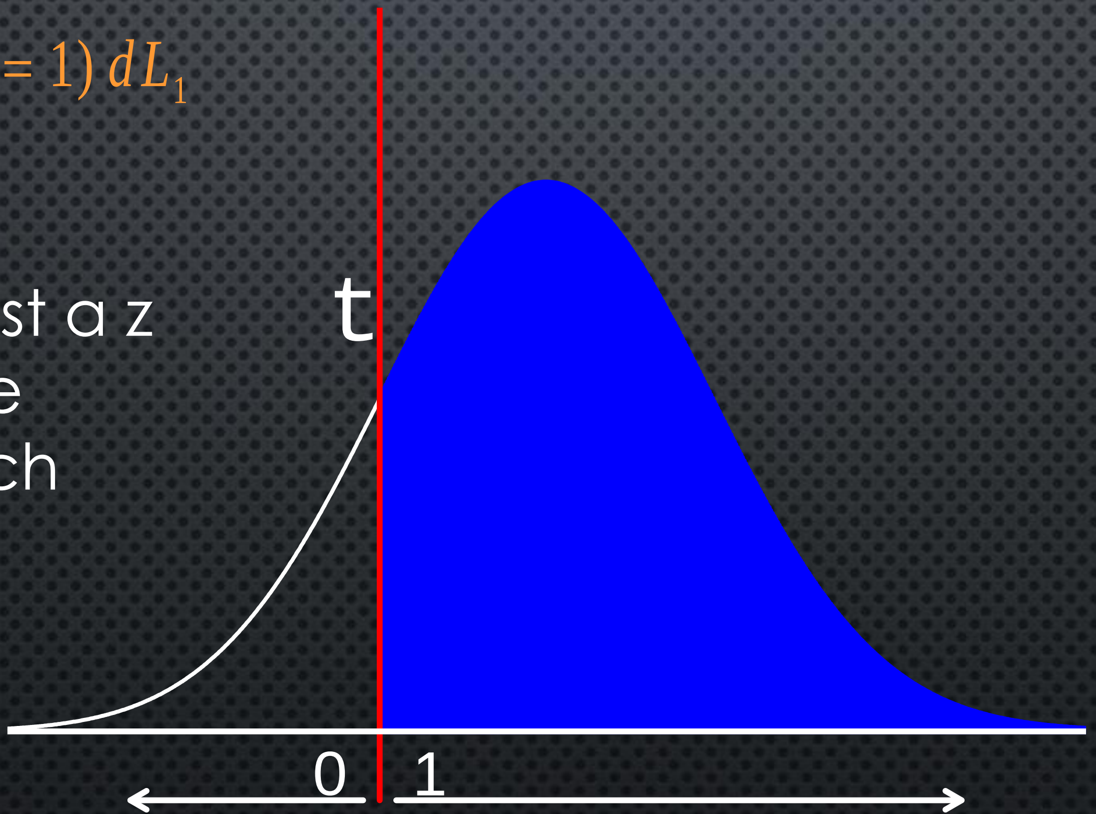
IDENTIFYING ASSUMPTIONS OF ORDINAL ASSOCIATIONS

- THE ASSUMPTIONS ARE ARBITRARY
  - THE SAME MODEL CAN BE SPECIFIED IN DIFFERENT WAYS, AND THE PARAMETERS WILL ESTIMATE DIFFERENT THINGS, BUT THE -2LNL SHOULD BE THE SAME FOR MODELS THAT ARE TRANSFORMATIONS OF EACH OTHER.
- THE ASSUMPTIONS ARE NECESSARY.
  - BECAUSE THE LATENT DIMENSION IS ONLY MEASURED INDIRECTLY, BY ORDINAL ITEMS, WE HAVE NO DIRECT INFORMATION ON ITS VARIANCE. THE THRESHOLDS COULD EXPAND OR CONTRACT (THINK ACCORDION) TO COMPLETELY COMPENSATE FOR A CHANGE IN VARIANCE.

INTUITIVE EXPLANATION OF THRESHOLDS IN THE UNIVARIATE NORMAL DISTRIBUTION

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

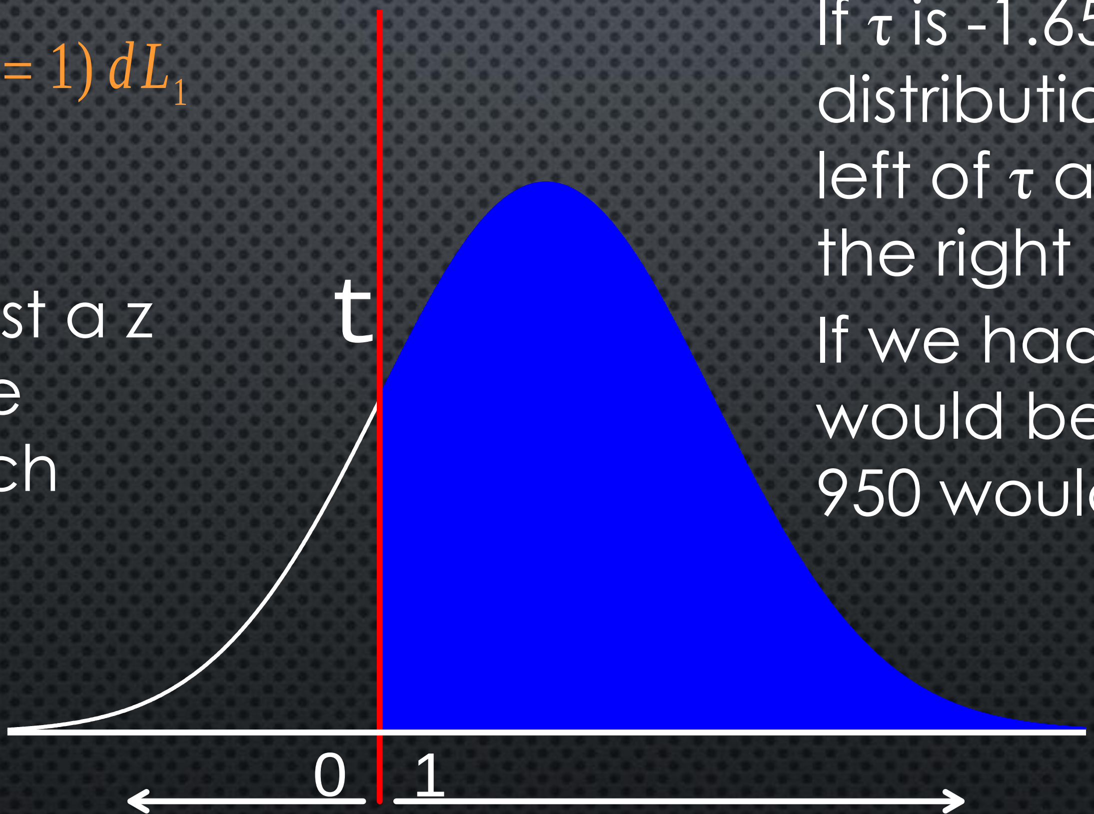
The threshold is just a z score and can be interpreted as such



INTUITIVE EXPLANATION OF THRESHOLDS IN THE UNIVARIATE NORMAL DISTRIBUTION

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

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



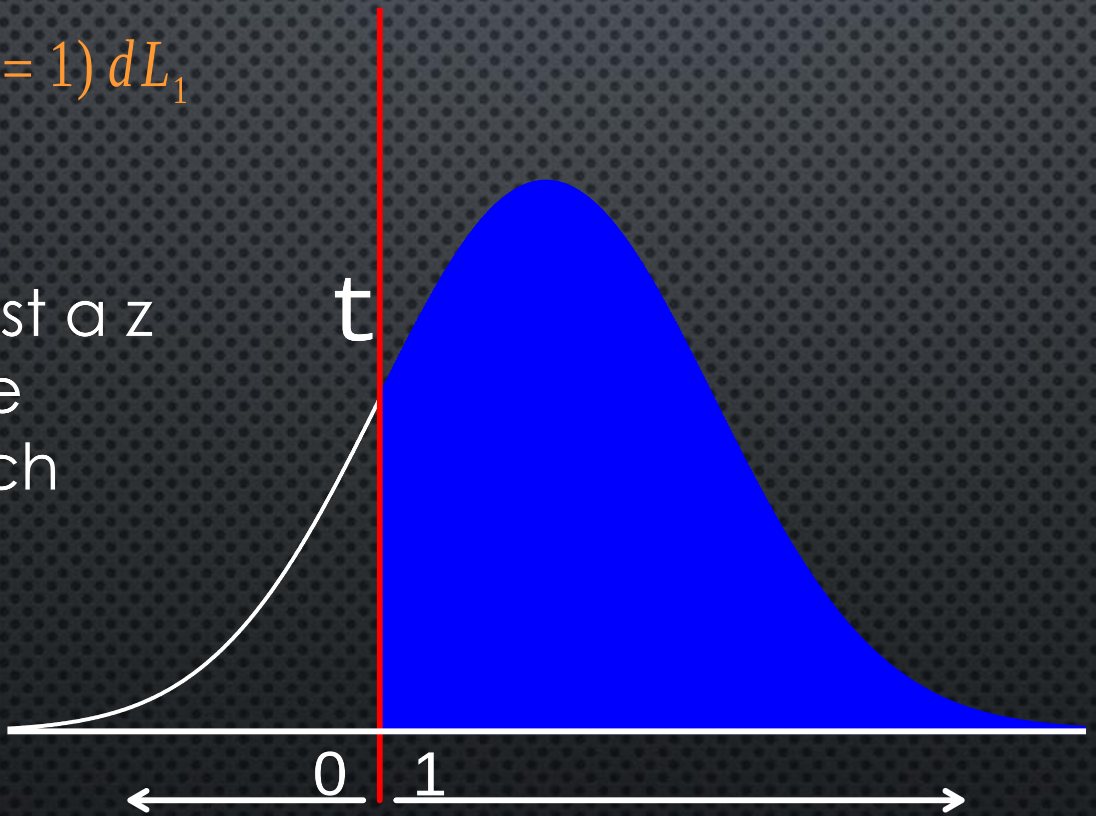
If τ is -1.65 then 5% of the distribution will be to the left of τ and 95% will be to the right

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

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

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



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

TWO BINARY TRAITS (E.G., DATA FROM TWINS)

Contingency Table with 4 observed cells:

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

Twin 1

Twin 2

	0	1
0	a	b
1	c	d

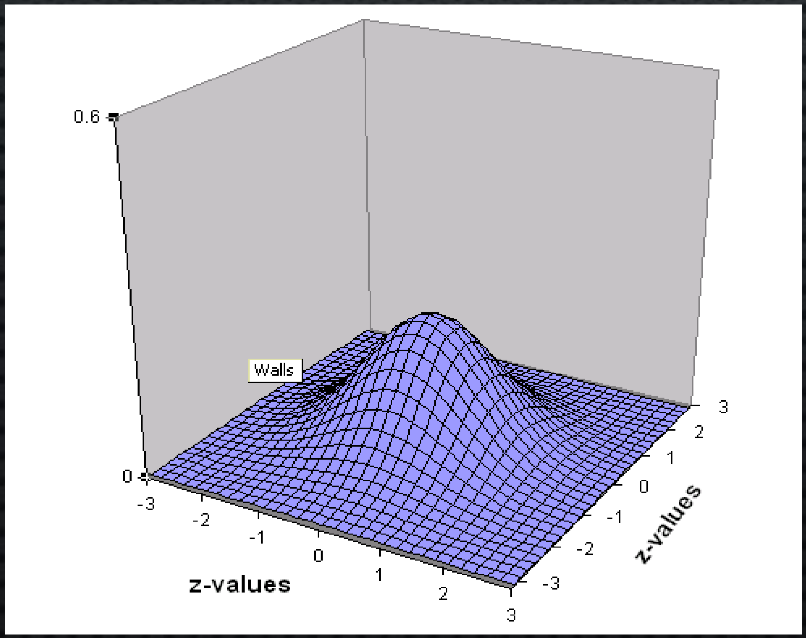
0 = unaffected
1 = affected

JOINT LIABILITY THRESHOLD MODEL FOR TWIN PAIRS

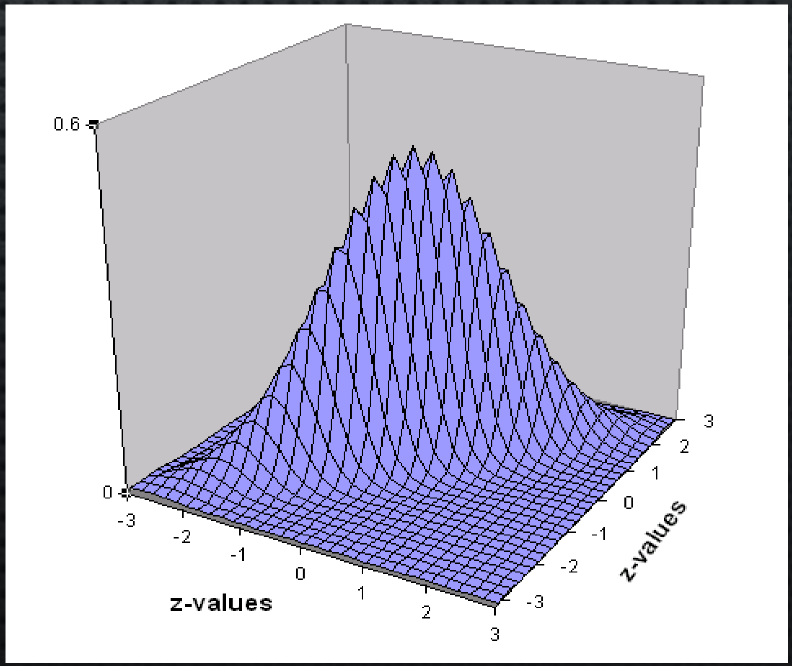
- Pairs are assumed to follow a **bivariate normal distribution**, where both traits have a mean of 0 and standard deviation of 1, and the **correlation** between them is what we want to know.

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

***r* =.00**

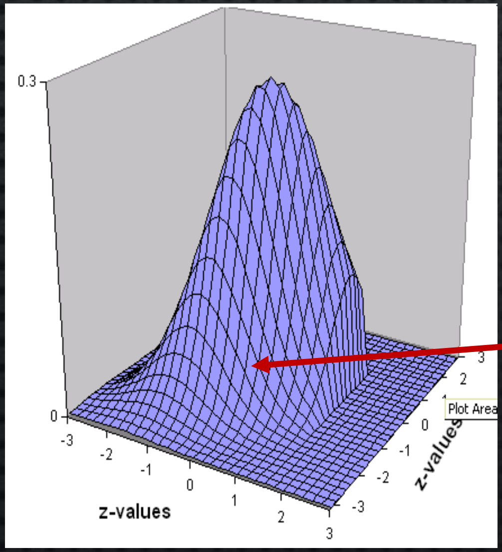


***r* =.90**

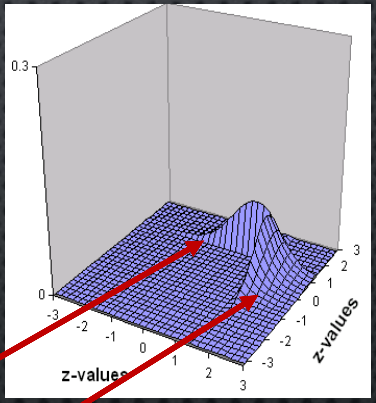


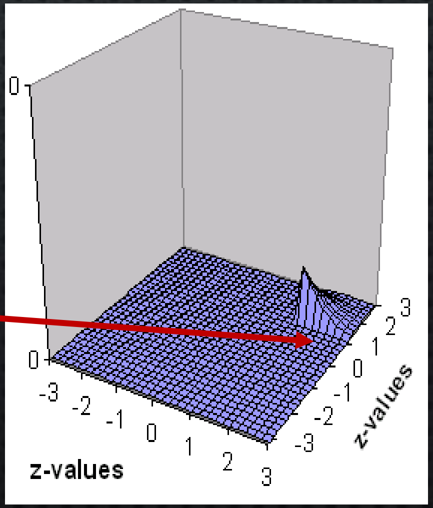
- The observed cell proportions relate to the proportions of the Bivariate Normal Distribution with a certain correlation between the latent variables (y_1 and y_2), each cut at a certain threshold

In other words, the joint probability of a certain response combination is the volume under the Bivariate Normal Distribution surface bounded by appropriate thresholds for each liability



y2 y1	Un	Aff
Unaffected	00	01
Affected	10	11

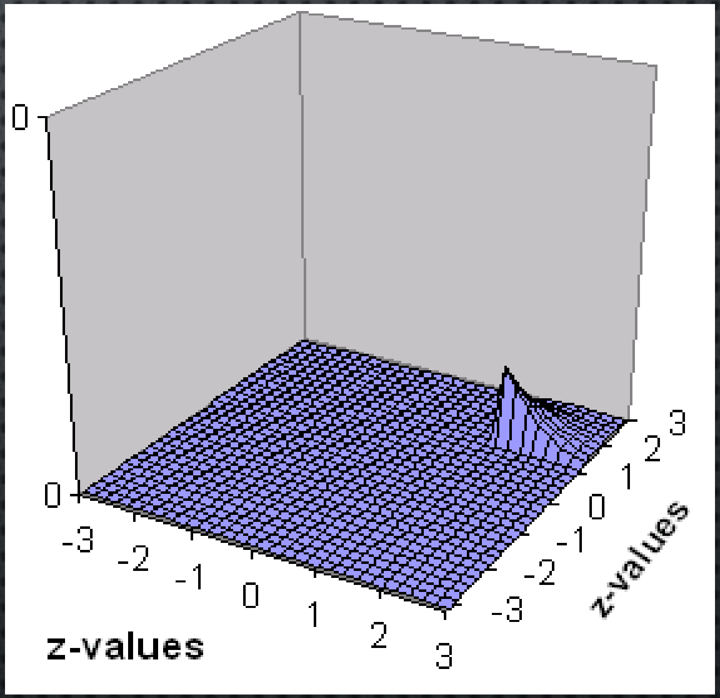




To calculate the cell proportions we rely on **Numerical Integration** of the Bivariate Normal Distribution over the two liabilities

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

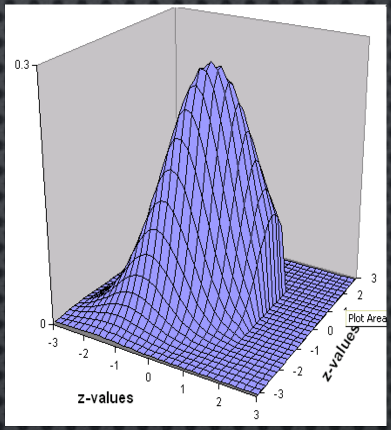
$$\int_{T_{c1}}^{\infty}\int_{T_{c2}}^{\infty}\Phi(y_1, y_2; \mu = 0, \Sigma)\,dy_1\,dy_2$$



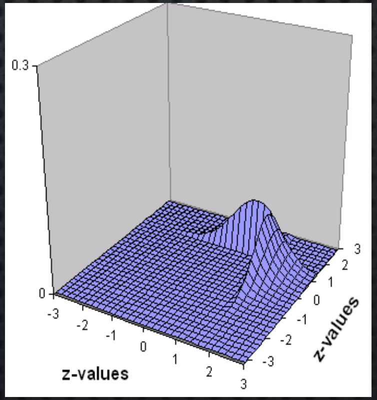
Φ is the bivariate normal probability density function,
y_1 and y_2 are the liabilities of twin1 and twin2,
with means of 0, and Σ the correlation between the two liabilities
T_{c1} is threshold (z-value) on y_1, T_{c2} is threshold (z-value) on y_2

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.

Y_2

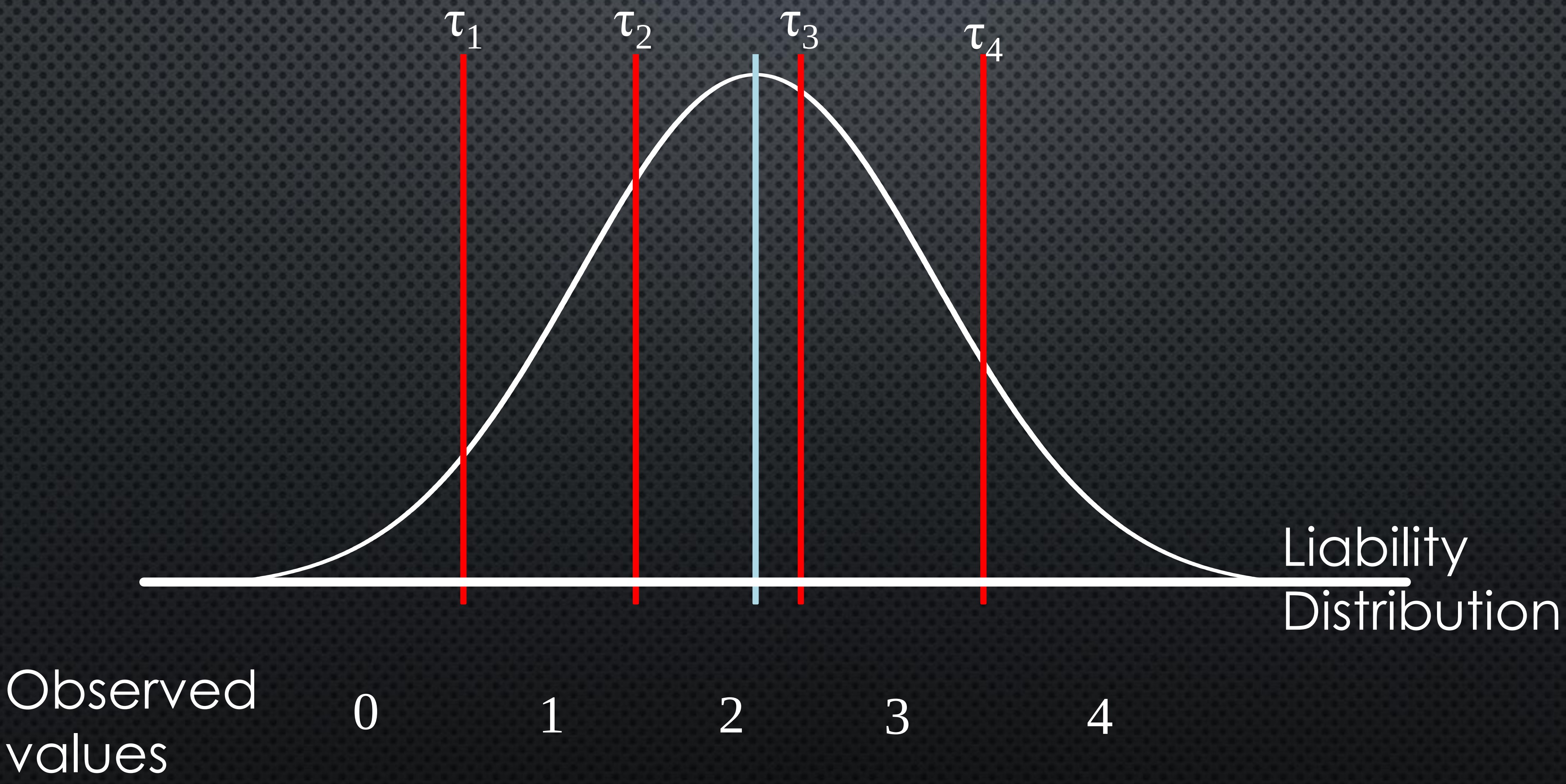
y1	0	1
0	.87	.05
1	.05	.03

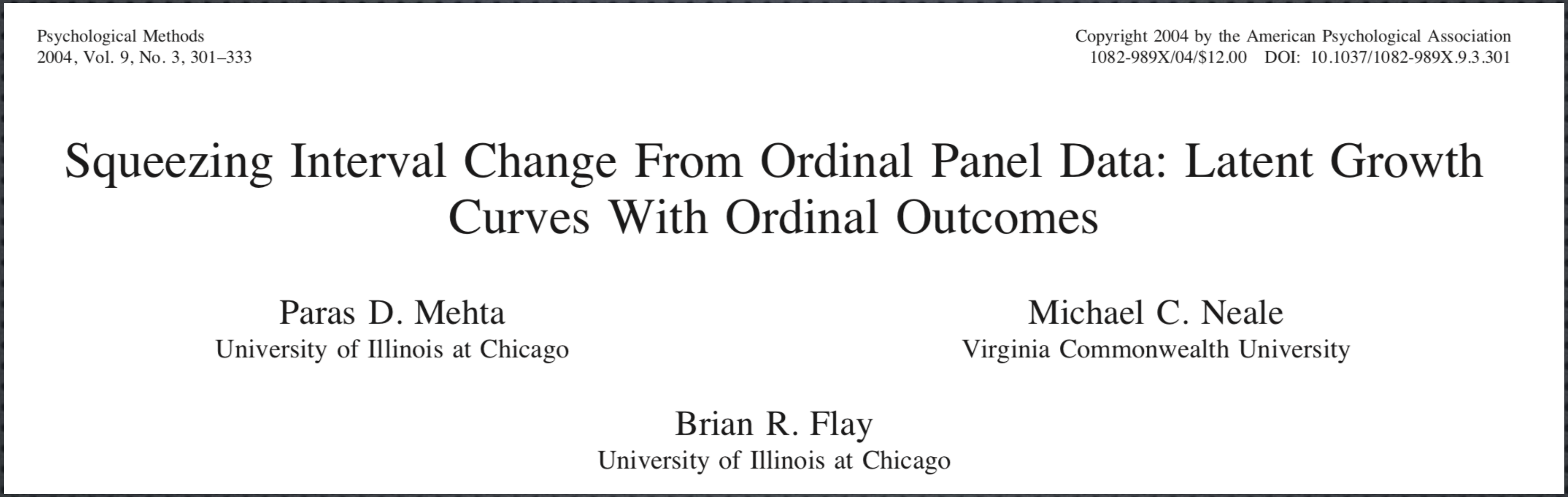
Y_1

$r = 0.60$

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

THE MULTIPLE THRESHOLD LIABILITY MODEL





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

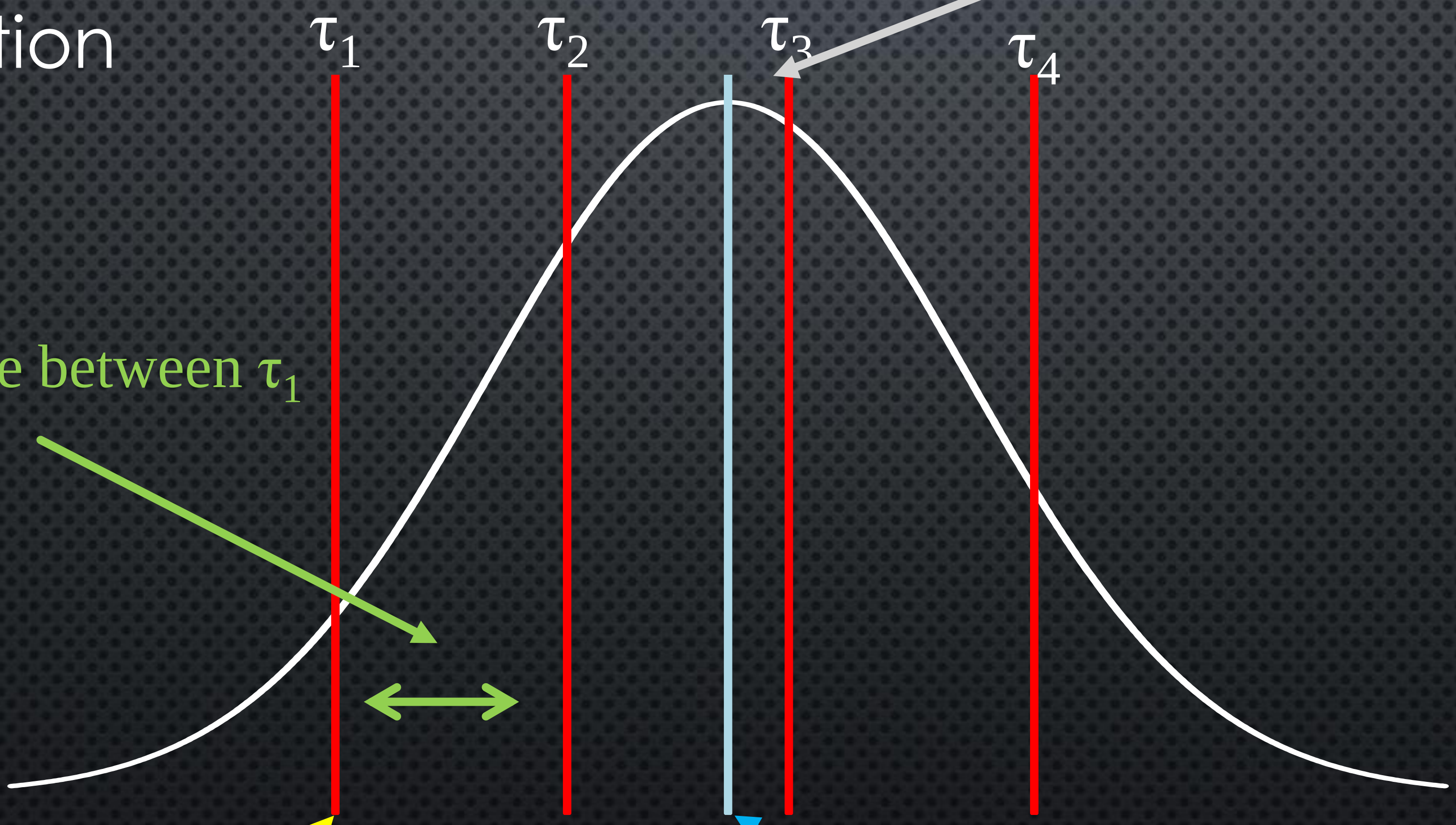
$\mathrm{Var}(\varepsilon | x) = 1$ in the normal-ogive model

Remember we can make different identifying assumptions but the model fit will stay the same

WHAT ALTERNATIVE ASSUMPTIONS COULD WE MAKE?

Mean of the distribution

Distribution of the liability



The distance between τ_1 and τ_2 is 1

τ_1 is is fixed to 0

The mean is freely estimated

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***

THANK YOU!