Welcome to Day 3!

To prepare your files for today please:

Create new directory in your workspace:

```
mkdir Day-3/
```

Enter the directory and copy files:

```
cd Day-3
cp /faculty/chelsea/2024/Day-3/Ordinal -R .
```

Do NOT forget the ‘.’ at the end of the above line!

WORKING WITH ORDINAL DATA

Chelsea Sawyers and Elizabeth Prom-Wormley

Objectives

- Explain why binary and ordinal data cannot be treated as continuous measures in analyses
- How to implement thresholds in OpenMx

Ordinal data

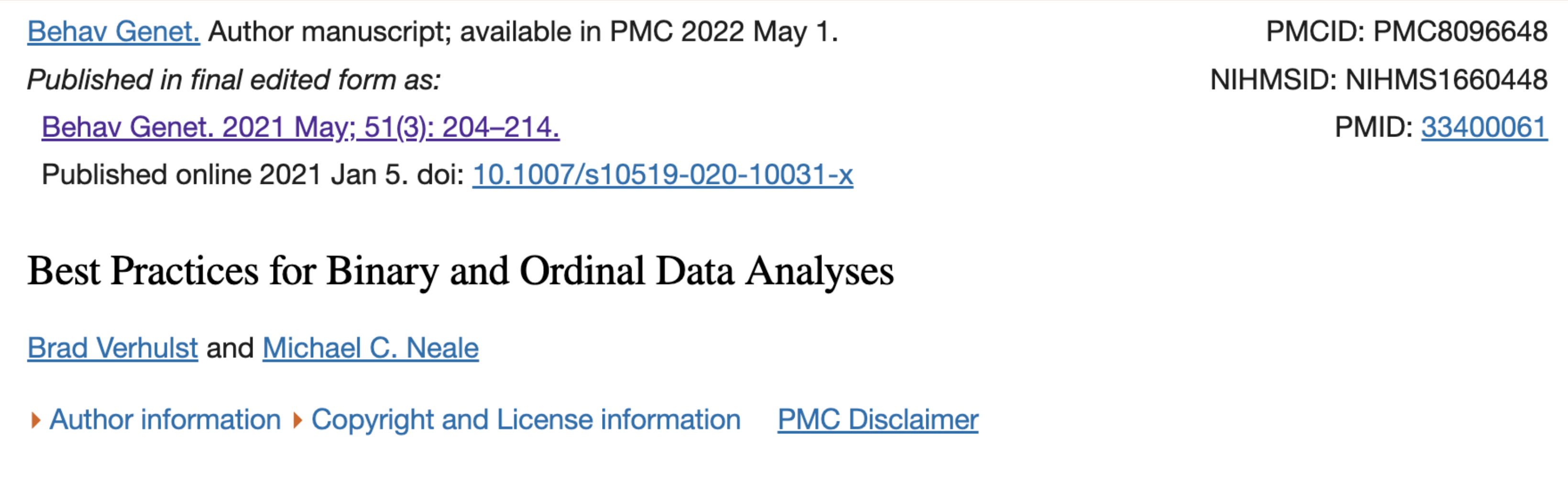
- Often measure behaviors using limited number of <u>ordered</u> categories:
  - *Absence (0) or presence (1) of a disorder*
  - *Severity of a disorder*
  - *Score on a single Likert item 'none/some/lots'*
  - *Number of symptoms (far from ideal)*

- In such cases the data take the form of counts, i.e. the number of individuals within each category of response

ORDINAL DATA REQUIRES DIFFERENT STATICAL METHODS

Using Continuous methods on Ordinal Data

- Underestimate the correlation/parameters
  - *Worse for binary than for categorical/ordinal*



Behav Genet. Author manuscript; available in PMC 2022 May 1.

PMCID: PMC8096648

Published in final edited form as:

NIHMSID: NIHMS1660448

Behav Genet. 2021 May; 51(3): 204–214.

PMID: 33400061

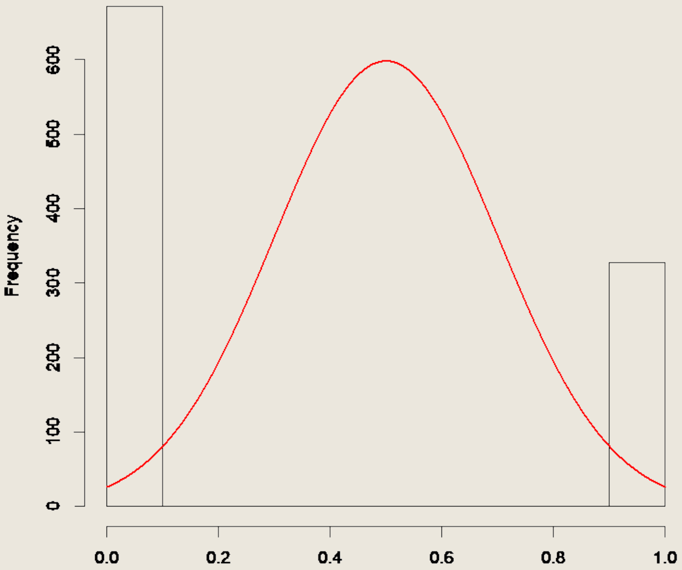
Published online 2021 Jan 5. doi: 10.1007/s10519-020-10031-x

Best Practices for Binary and Ordinal Data Analyses

Brad Verhulst and Michael C. Neale

▸ Author information ▸ Copyright and License information PMC Disclaimer

Problems with the treating ordinal variables as continuous



- Normality – Ordinal 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*

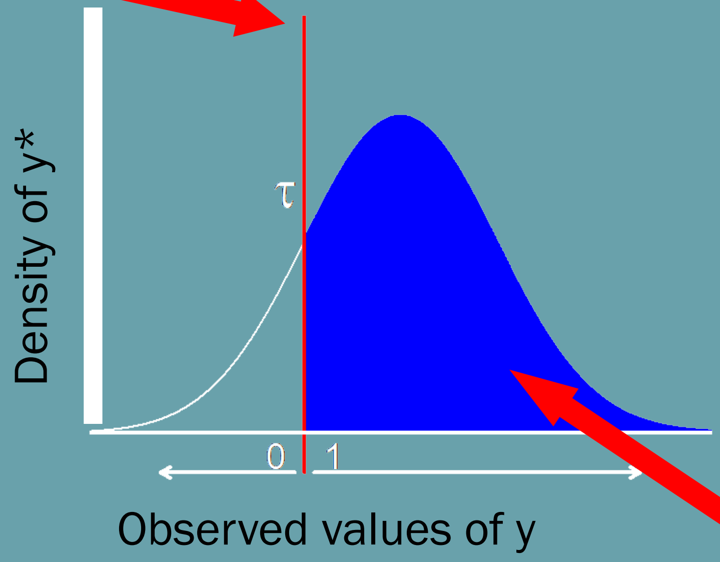
2. Assume the Binary Variable as a random draw from a Binomial (or Bernouilli) Distribution (Non-Linear Probability Model). Genuinely categorical responses, no underlying continuous distribution.

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 Threshold Model

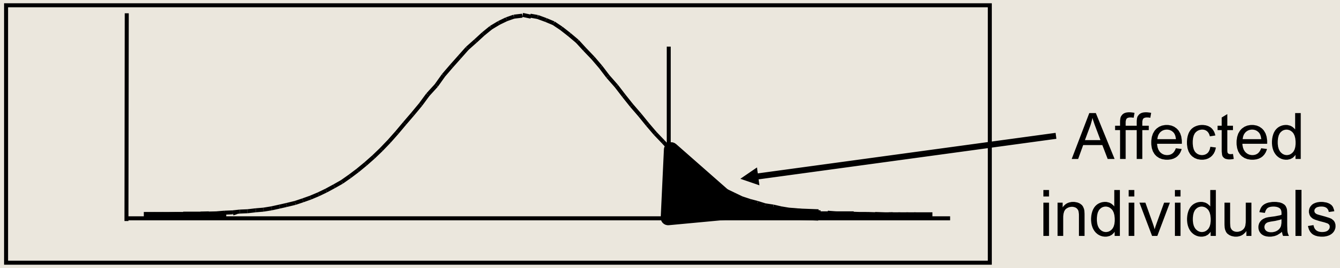
Threshold



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

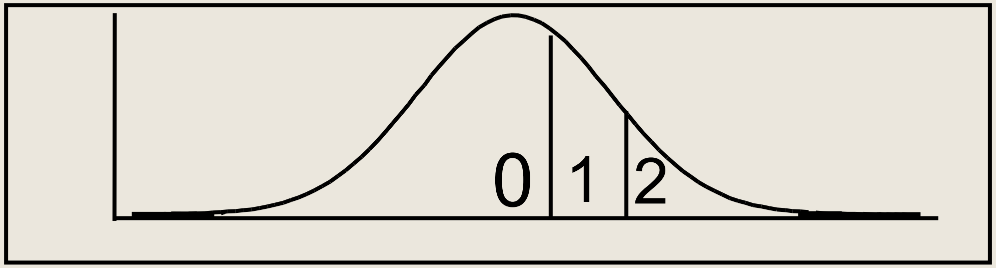
Liability Dimension

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

Ideas 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

Mean Assumption

The intercept (mean) is 0

The traditional assumption

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

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

The Probit Model

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

The Logit Model

Assumption 3

The conditional mean of ε is 0.

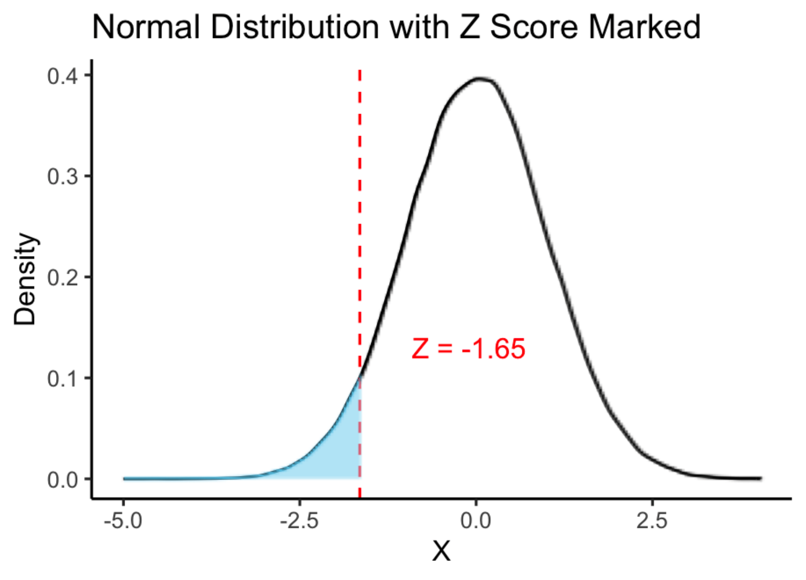
- *This is the same assumption as we make for continuous variables, and allows the parameters to be unbiased*

Identifying Assumptions of Ordinal Associations

- The assumptions are arbitrary
  - *The same model can be specified in different ways, but the parameters will estimate different things. 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.*

Fundamentals of the Threshold Model

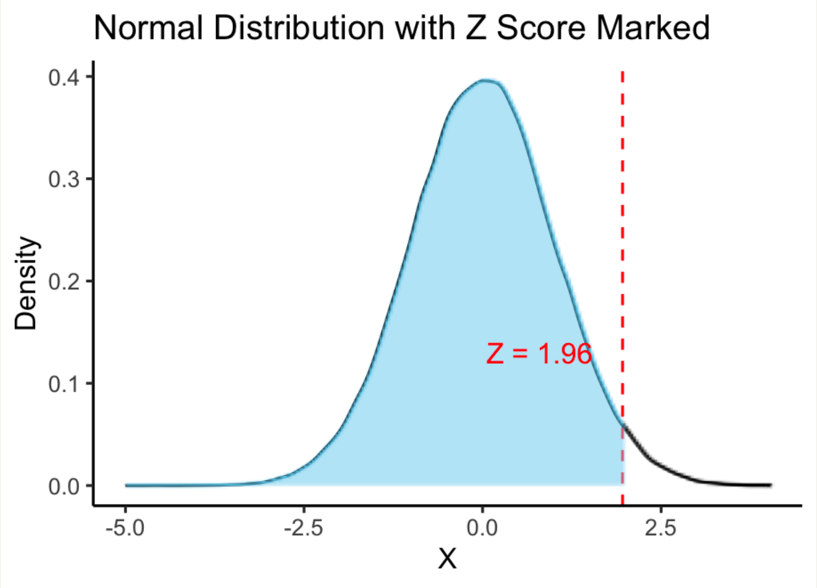


- If τ is -1.65 then 5% of the distribution will be to the left of τ and 95% will be to the right

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

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

Fundamentals of the Threshold Model



- If τ is 1.96 then 97.5% of the distribution will be to the left of τ and 2.5% will be to the right

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

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

TIME FOR DATA

Open BinaryWarmup.R

Copy from: /home/chelsea/2024/Day-3/Ordinal
If you did not do so at the beginning

WHAT ABOUT TWINS?

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

	0	1
0	a	b
1	c	d

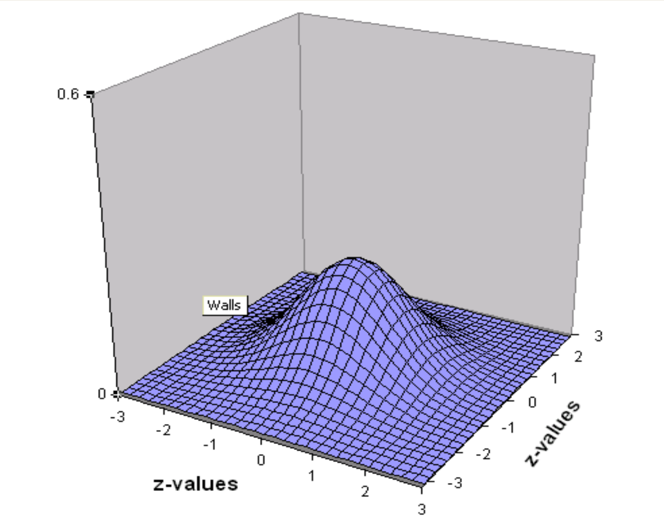
0 = unaffected
1 = affected

Joint Liability Threshold Model for twin pairs

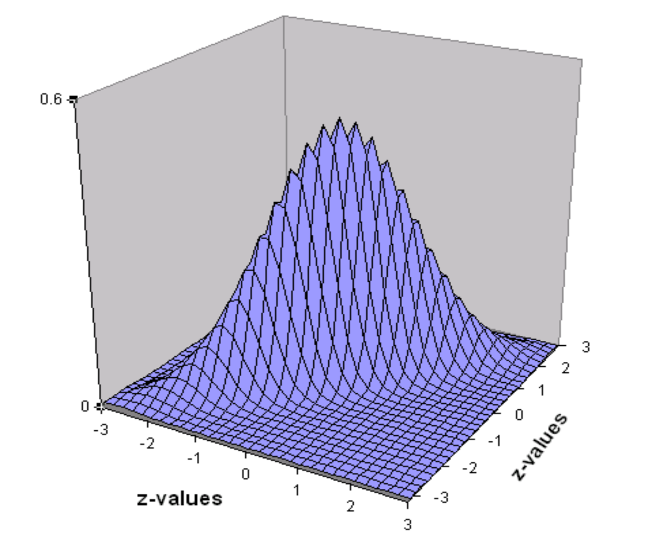
- Pairs are assumed to follow a **bivariate normal** distribution, where both traits have a mean of 0 and standard deviation of 1, and the **correlation** between them is what we want to know.

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

***r* =.00**

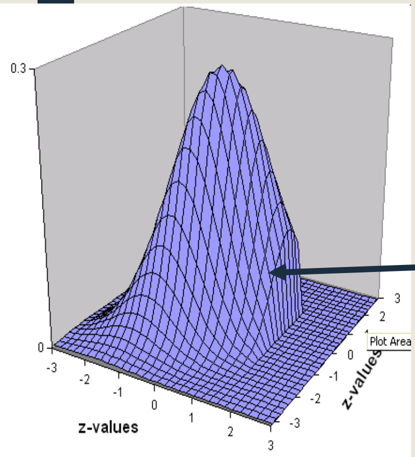


***r* =.90**

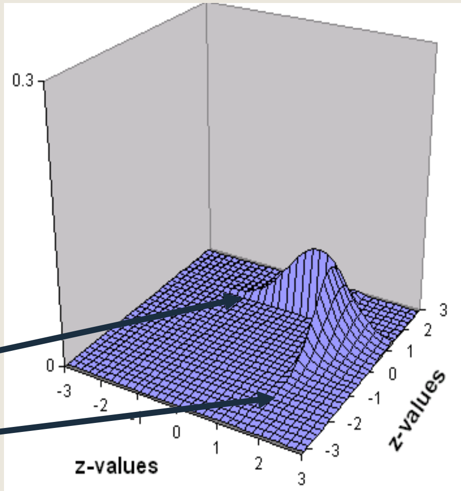


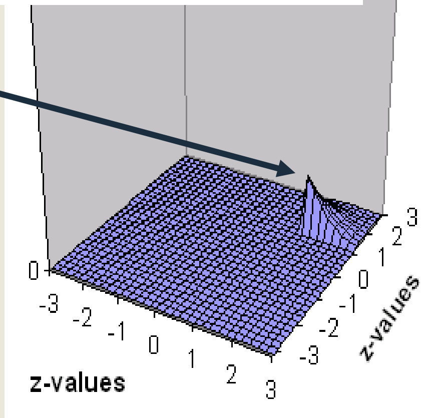
- The observed cell proportions relate to the proportions of the Bivariate Normal Distribution with a certain correlation between the latent variables (y_1 and y_2), each cut at a certain threshold

In other words, the joint probability of a certain response combination is the volume under the BND surface bounded by appropriate thresholds on each liability



y2 y1	0	1
0	100	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

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.

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

$$r = 0.60$$

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

More Practice Available!

If you would like to work more with this threshold model to see how correlations shape the bivariate normal distribution and how we can move from correlations to estimated cell proportions in contingency tables, please work through the Ordinal Data Challenge R script and Qualtrics on your own time

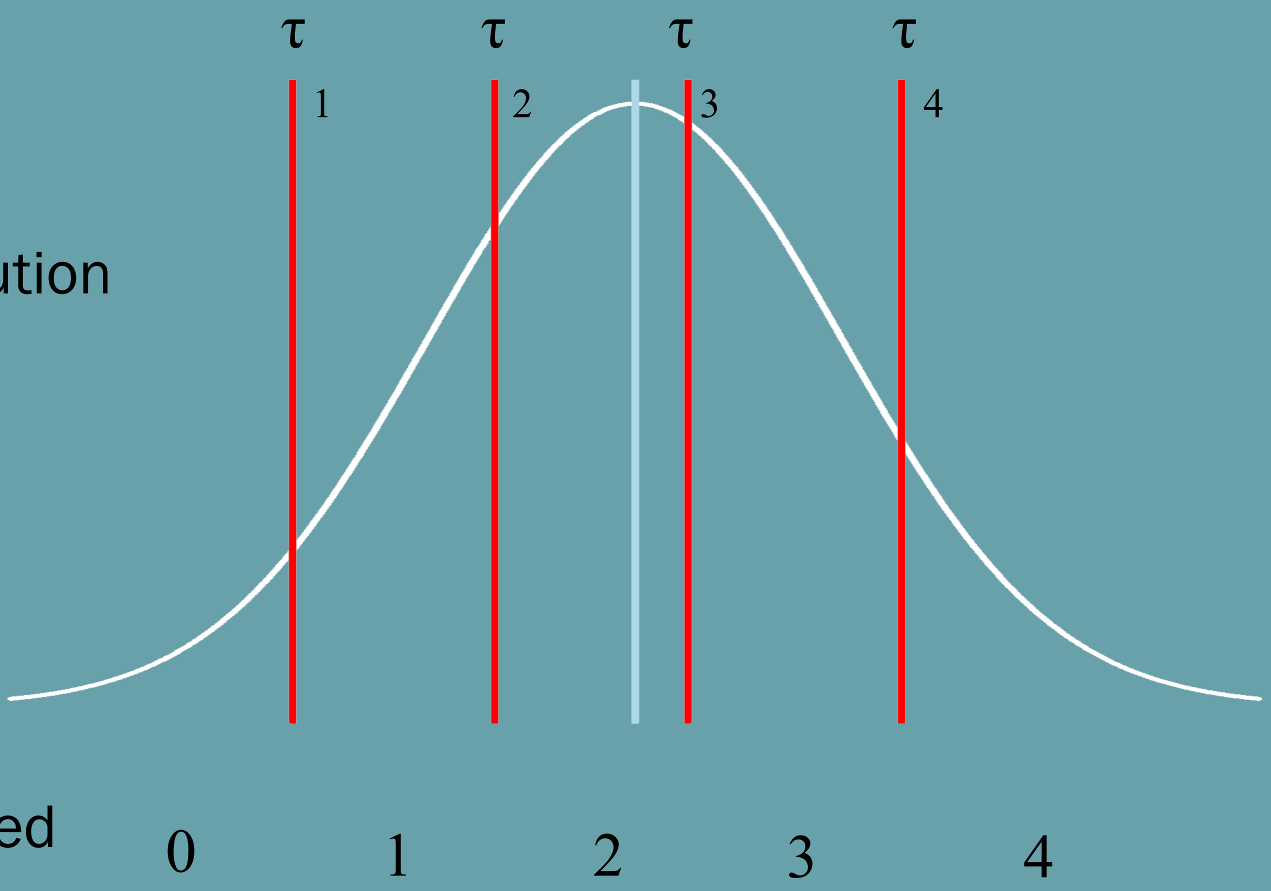
https://qimr.az1.qualtrics.com/jfe/form/SV_4YoLFEuPnBc6QCy

MOVING TO ORDINAL DATA!

Intuition behind the Multiple Threshold Liability model

Distribution of the liability

Observed values



NOW TIME FOR TWINS!

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***

Transitioning from Continuous Logic to Ordinal Logic

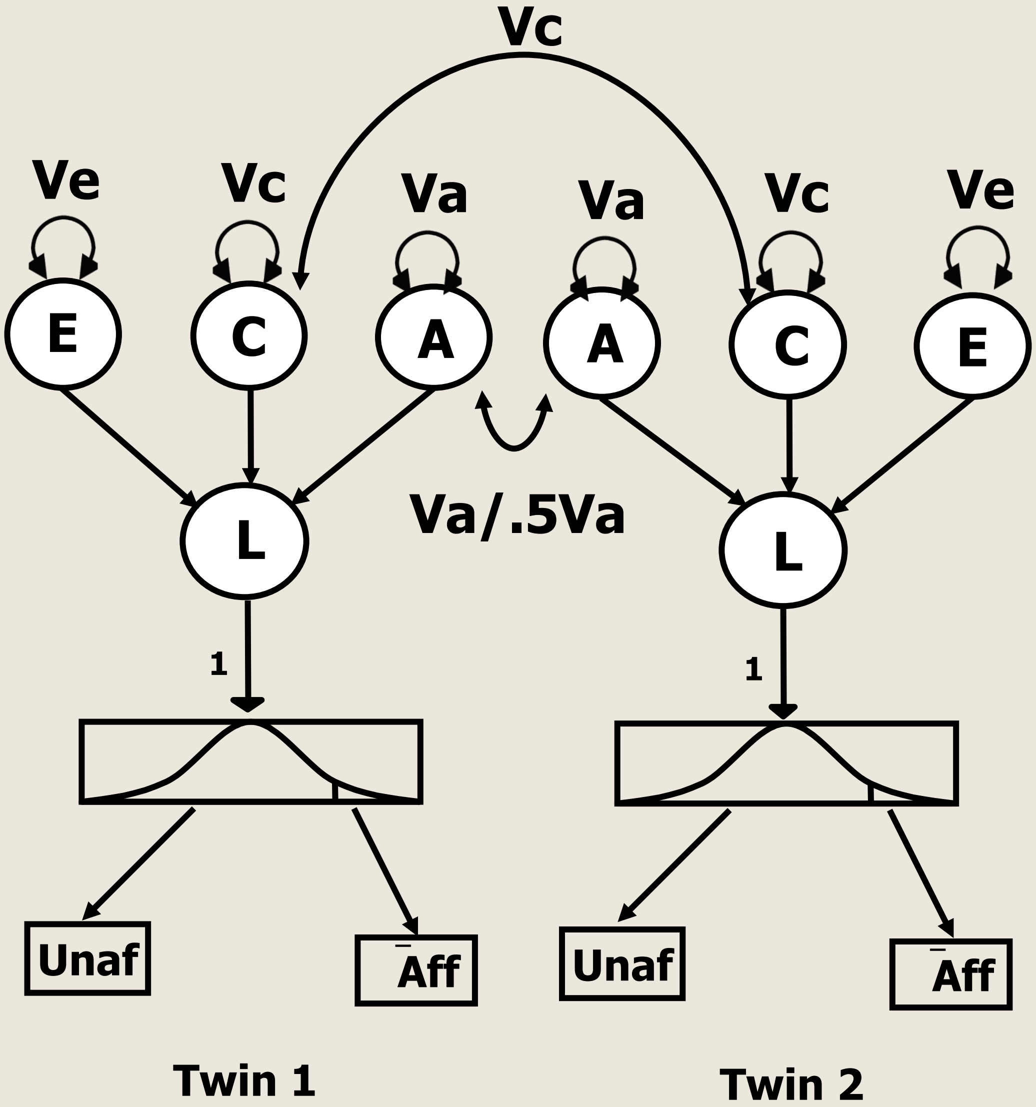
- Ordinal data has 1 less degree of freedom compared to continuous data
  - MZcov, DZcov, Prevalence
  - No information on the variance
- Thinking about our ACE/ADE model
  - 4 parameters being estimated
  - A C E mean
- ACE/ADE model is unidentified without adding a constraint

Two Approaches to the Liability Threshold Model

- Traditional
  - *Maps data to a standard normal distribution*
  - *Total variance constrained to be 1*

- Alternate
  - *Fixes an alternate parameter (usually E)*
  - *Estimates the remaining parameters*

ACE Liability Model



Variance constraint

Threshold model

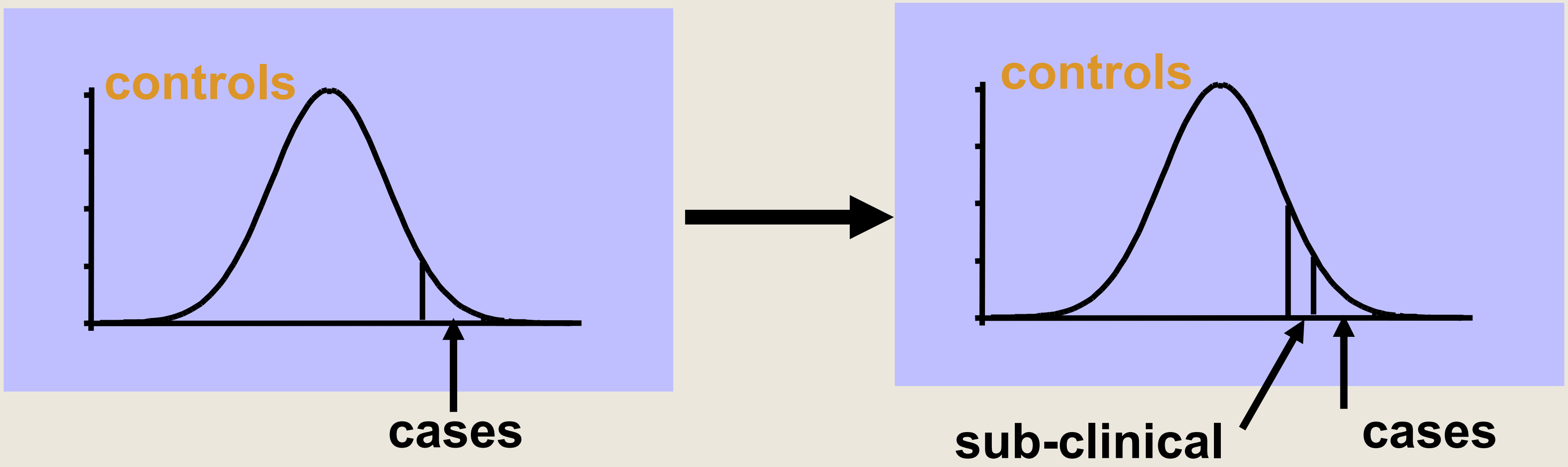
Power issues

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

Neale, Eaves & Kendler 1994

- Solutions:

1. Bigger samples
2. Use more categories



TIME FOR MORE DATA

Open oneACEvo.R

Copy from: /faculty/chelsea/2024/Day-3/Ordinal
If you did not do so at the beginning

TRANSLATING BACK TO THE SEM APPROACH IN OPENMX

Univariate Analysis with Ordinal Data
A Roadmap

1*- Use the data to test basic assumptions inherent to standard ACE (ADE) models

Saturated Model

2- Estimate contributions of genetic and environmental effects on the liability of a trait

ADE or ACE Models

3**- Test ADE (ACE) submodels to identify and report significant genetic and environmental contributions

AE or E Only Models

**Not reviewing saturated model in this practical*

*** Maybe if there's time*

BACK TO R…

Copy from: /faculty/chelsea/2024/Day-3/Ordinal
If you did not do so at the beginning

Handling Ordinal Data in OpenMx

- # Declare variables to be ordered Factors for OpenMx

```
mzDataF <- mxFactor( x=mzData, levels=c(0:nth) )
dzDataF <- mxFactor( x=dzData, levels=c(0:nth) )
```

Handling Ordinal Data in OpenMx

- 1- Determine the 1st threshold

```
svLTh <- -1.5          # start value for first threshold
```

- 2- Determine displacements between 1st threshold and subsequent thresholds

```
svITh <- 1             # start value for increments
```

- 3- Add the 1st threshold and the displacement to obtain the subsequent thresholds

```
svTh <- matrix(rep(c(svLTh,(rep(svITh,nth-1)))),nrow=nth,ncol=nv)
```

-1.5
1
1

svTh = (1x nth)

Matrix and Algebra for Expected Means

```
meanG    <- mxMatrix( type="Zero", nrow=1, ncol=nv, name=”meanG" )
```

0	0

1X2 Matrix

Matrices for Expected Thresholds

```
thinG <- mxMatrix( type="Full", nrow=nth, ncol=ntv,
    free=TRUE, values=svTh, lbound=lbTh,
    labels=labTh("th",vars,nth), name="thinG" )
```

Th1	**th1**
inc1	**inc1**
Inc1	**inc1**

3X2 Matrix

* The positive bounds on the increments stop the thresholds going ‘backwards’, i.e. they preserve the ordering of the categories

```
Inc    <- mxMatrix( type="Lower", nrow=nth,
    ncol=nth, free=FALSE, values=1, name="Inc" )
```

1	0	0
1	1	0
1	1	1

3X3 Matrix

Algebra for Expected Thresholds

```
threG   <- mxAlgebra( expression= inc %*% thinG,
    name="threG" )
```

1	0	0
1	1	0
1	1	1

% * %

Th1	th1
inc1	inc1
Inc1	inc1

=

Th1	Th1
Th1 +inc	Th1 +inc
Th1 +inc2	Th1 +inc2

A multiplication is used to ensure that any threshold is higher than the previous one. This is necessary for the optimization procedure involving numerical integration over the MVN

Note: this only works if the increments are POSITIVE values, therefore a BOUND statement around the increments are necessary

With labels=

t1thobmi	t1thobmi
t2thobmi	t2thobmi
t3thobmi	t3thobmi

ADE Model Deconstructed
Variance Components

```
covA <- mxMatrix( type="Symm", nrow=nv, ncol=nv,
free=TRUE, values=svPa, label="VA11", name="VA" )
```

VA

1 x 1 matrix

```
covD <- mxMatrix( type="Symm", nrow=nv, ncol=nv,
free=TRUE, values=svPa, label="VD11", name="VD" )
```

VD

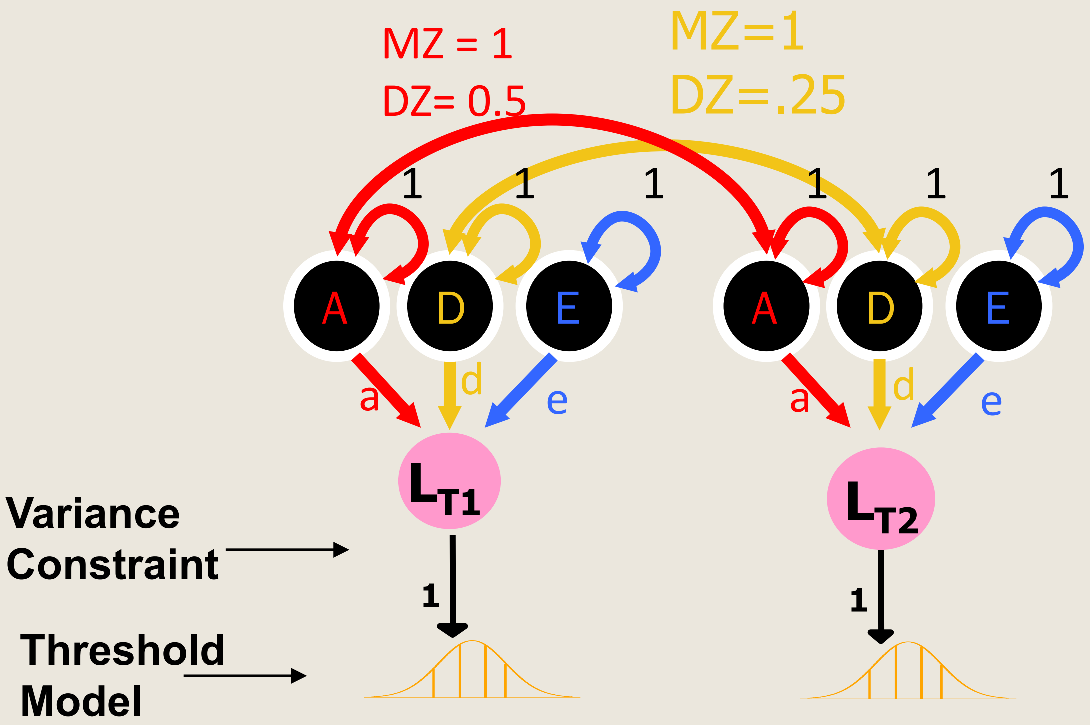
1 x 1 matrix

```
covE <- mxMatrix( type="Symm", nrow=nv, ncol=nv,
free=TRUE, values=svPa, label="VE11", name=" VE" ) #
```

VE

1 x 1 matrix

ADE Liability Model



```
covMZ    <- mxAlgebra(expression= VA+ 0.25%x%VD, name="cMZ" )
covDZ    <- mxAlgebra( expression= 0.5%x%VA+ 0.25%x%VD, name="cDZ”)

expCovMZ <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)),
name="expCovMZ" )
expCovDZ <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)),
name="expCovDZ" )
```

ADE Model Deconstructed

Constraint on Variance of Ordinal Variables

```
covP     <- mxAlgebra( expression= VA+VD+VE, name=" V " )
```

$\boxed{V}$

1 x 1 matrix

```
var1     <- mxConstraint(expression=diag2vec(V)==1,
name="Var1" )
```

$\boxed{1}$

1 x 1 matrix

Why is the variance being constrained?

LET’S LOOK AT THE OUTPUT

How Many Parameters?

- A
- D
- E

- Thresholds (3)

Questions to Consider

- Are there any submodels that are appropriate to use instead of ADE?
- What are your conclusions regarding genetic and environmental influences on BMI?

Objectives

- Explain why binary and ordinal data cannot be treated as continuous measures in analyses
- How to implement thresholds in OpenMx models

Resources for twin modeling

If you need twin code an excellent resource is:

https://hermine-maes.squarespace.com/#/onea5/

If you have OpenMx questions:

https://openmx.ssri.psu.edu/forums

Want to know more about OpenMx?

https://openmx.ssri.psu.edu/