Threshold Liability Models
(Ordinal Data Analysis)

Frühling Rijsdijk

MRC SGDP Centre, Institute of Psychiatry, King's College London

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 Q item e.g. : 0 - 1, 0 - 4

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

Analysis of ordinal variables

- The session aims to show how we estimate correlations from count data (with the ultimate goal to estimate h^2, c^2, e^2)
- 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

Liability

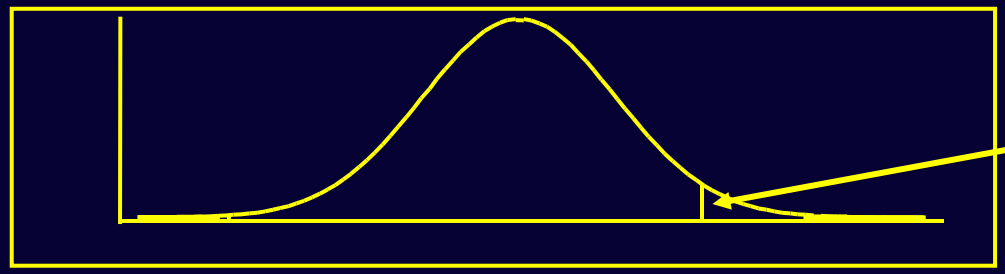
Liability is a **theoretical** construct. It's the assumption we make about the distribution of a variable which we were only able to measure in terms of a few ordered categories

Assumptions:

(1) Categories reflect an imprecise measurement of an underlying *normal distribution* of liability

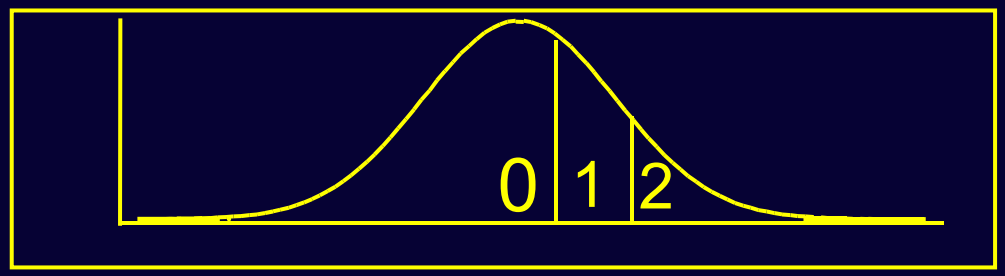
(2) The liability distribution has 1 or more thresholds (cut-offs) to discriminate between the categories

For disorders:



The ***risk*** or liability to a disorder is normally distributed, only when a certain threshold is exceeded will someone have the disorder. Prevalence: proportion of affected individuals.

For a single questionnaire item score e.g:



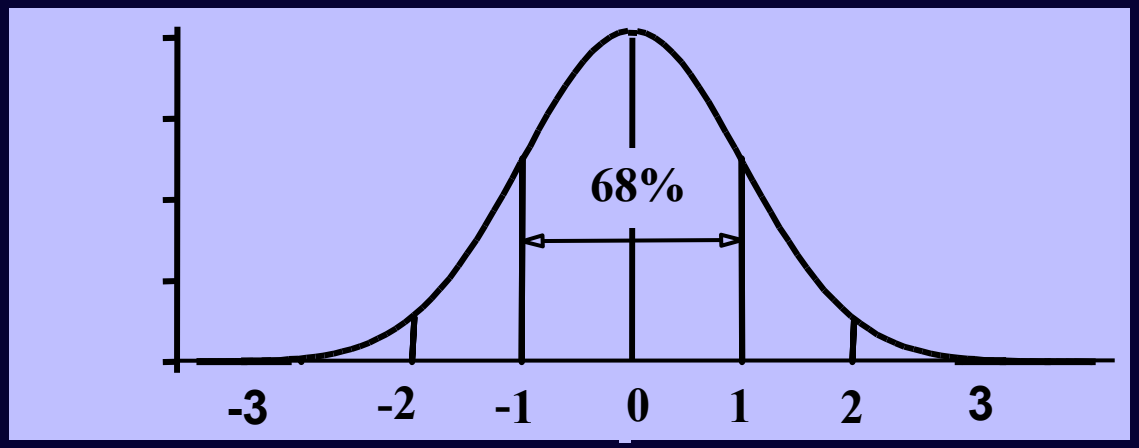
0 = not at all
1 = sometimes
2 = always

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

The Standard Normal Distribution

Liability is a *latent* variable, the scale is arbitrary, distribution is assumed to be a *Standard Normal Distribution* (SND) or z-distribution:

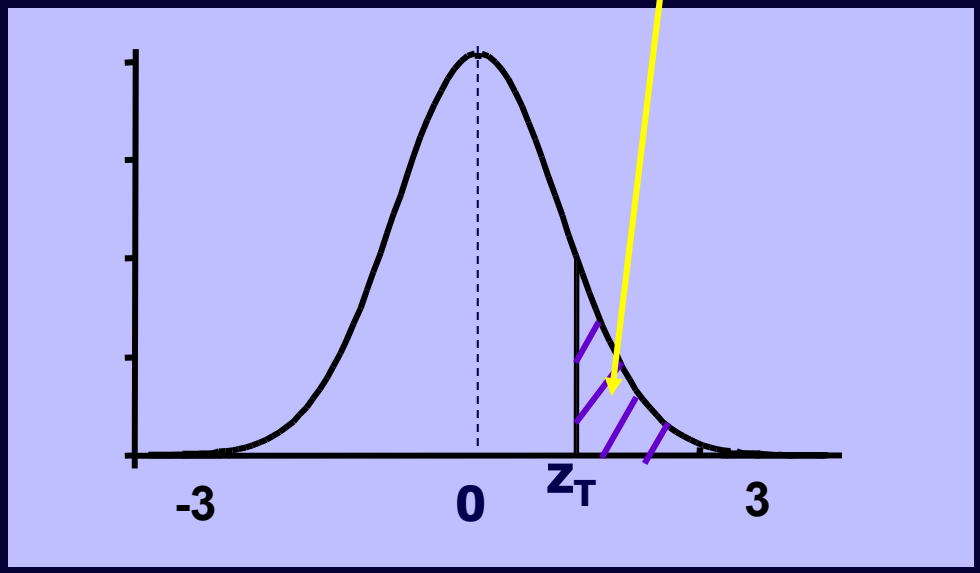
- Mathematically described by the SN Probability Density function (Φ =phi), a bell-shaped curve with:
  - mean = 0 and SD = 1
  - z-values are the number of SD away from the mean
- Convenience: area under curve =1, translates directly to probabilities



Standard Normal Cumulative Probability in right-hand tail

(For negative z values, areas are found by symmetry)

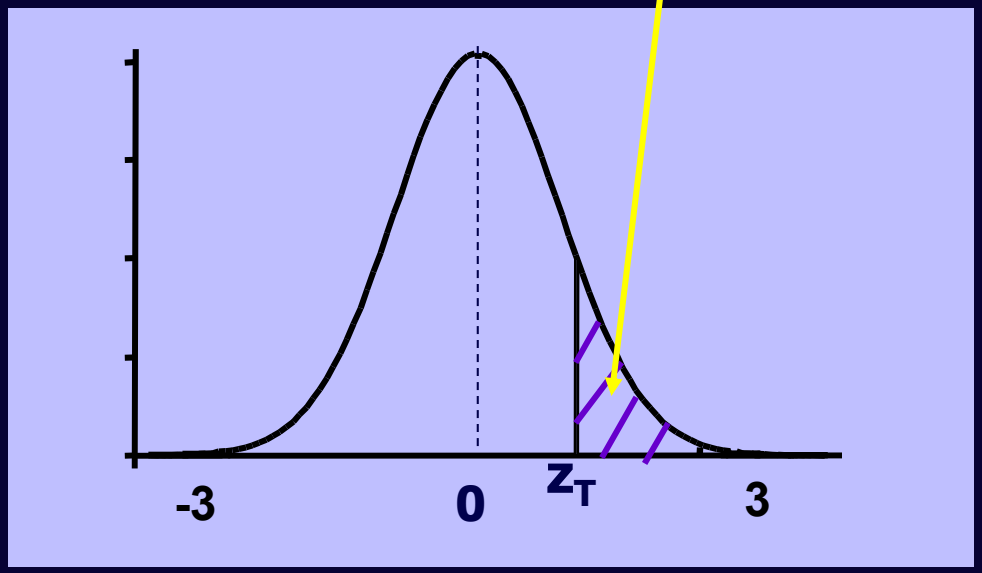
$$\text{Area} = P(z \geq z_T)$$



Standard Normal Cumulative Probability in right-hand tail

(For negative z values, areas are found by symmetry)

Area = P($z \geq z_T$)



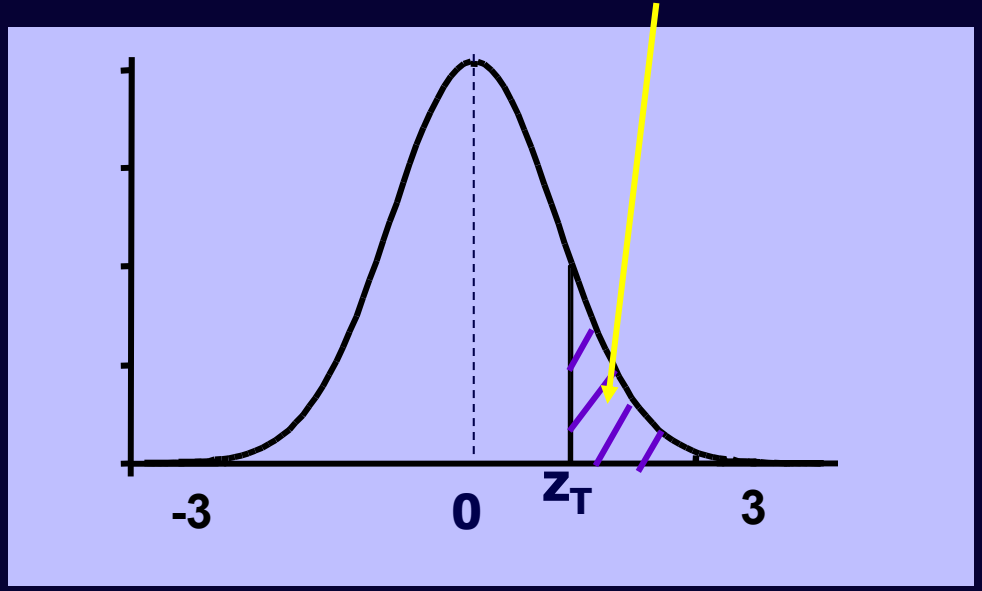
$$\int_{z_T}^{\infty} \Phi(L_1; \mu = 0, \sigma^2 = 1)\, dL_1$$

Standard Normal Cumulative Probability in right-hand tail

(For negative z values, areas are found by symmetry)

Z_0	Area	
0	.50	50%
.2	.42	42%
.4	.35	35%
.6	.27	27%
.8	.21	21%
1	.16	16%
1.2	.12	12%
1.4	.08	8%
1.6	.06	6%
1.8	.036	3.6%
2	.023	2.3%
2.2	.014	1.4%
2.4	.008	.8%
2.6	.005	.5%
2.8	.003	.3%
2.9	.002	.2%

Area = P($z \geq z_T$)



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

Two ordinal traits: Data from twins

> Contingency Table with 4 observed cells:

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

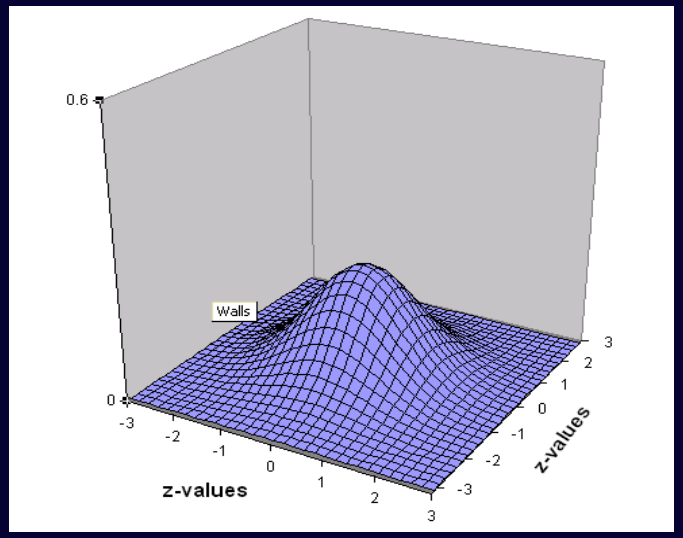
Twin1 / Twin2	0	1
0	a	b
1	c	d

0 = unaffected
1 = affected

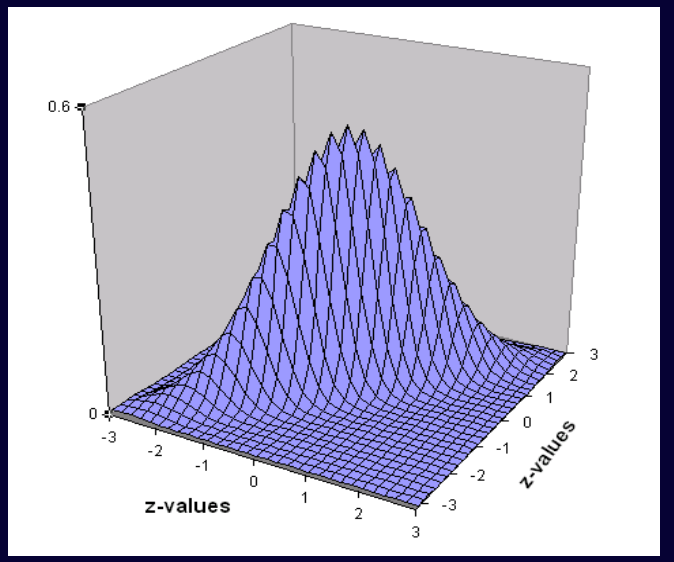
Joint Liability Model for twin pairs

- Assumed to follow a **bivariate normal** distribution, where both traits have a mean of 0 and standard deviation of 1, but the **correlation** between them is variable.
- The **shape** of a bivariate normal distribution is determined by the **correlation** between the traits

***r* =.00**

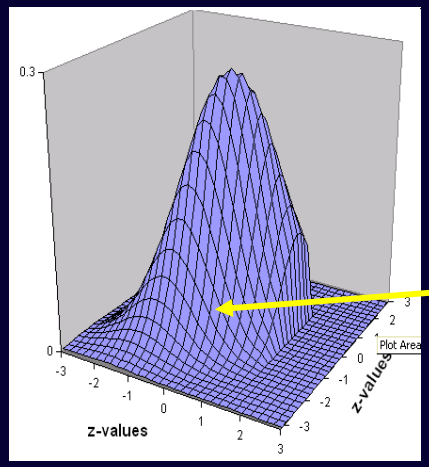


***r* =.90**

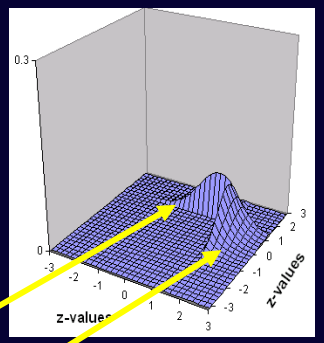


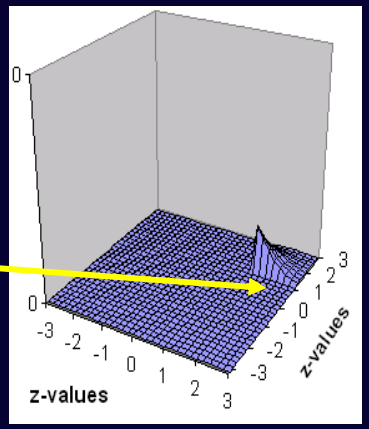
- The observed cell proportions relate to the proportions of the BND with a certain correlation between the latent variables (y_1 and y_2), each cut at a certain threshold

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



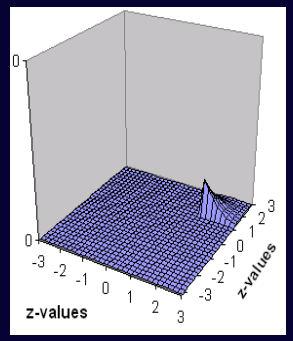
y1 \ y2	0	1
0	00	01
1	10	11





Expected cell proportions

Numerical integration of the BND over the two liabilities
e.g. the probability that both twins are above T*c* :

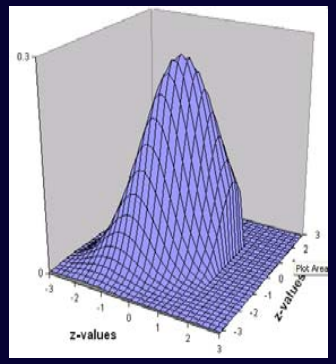


$$\int_{T_{c1}}^{\infty}\int_{T_{c2}}^{\infty}\Phi(y_1, y_2; \mu = 0, \Sigma)\,dy_1\,dy_2$$

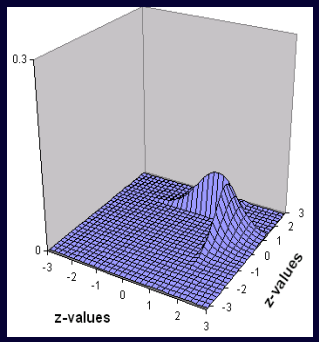
Φ is the bivariate normal probability density function,
y_1 and y_2 are the liabilities of twin1 and twin2,
with means of 0, and Σ the correlation between the two liabilities
T_{c1} is threshold (z-value) on y_1, T_{c2} is threshold (z-value) on y_2

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 BN 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)

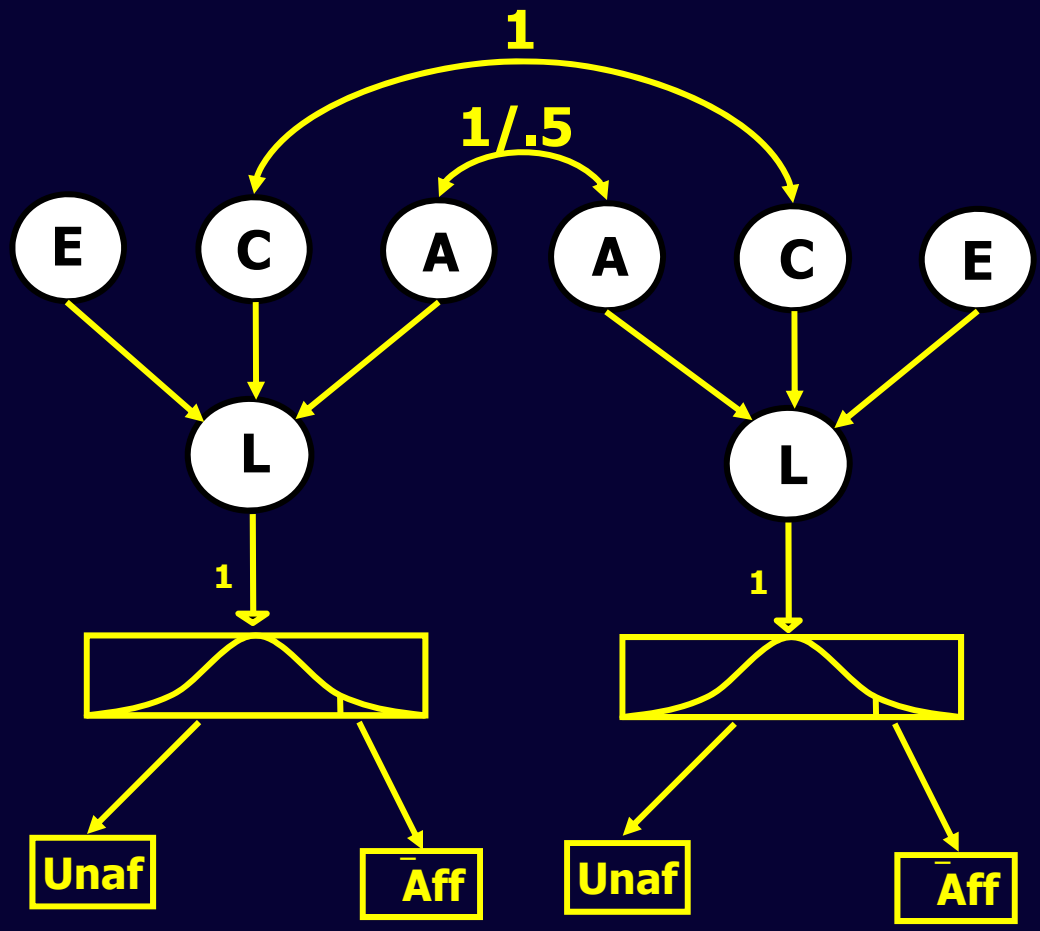
Bivariate Ordinal Likelihood

- The likelihood for each observed ordinal response pattern is computed by the expected proportion in the corresponding cell of the BN distribution
- The maximum-likelihood equation for the whole sample is -2* log of of the likelihood of each vector of observation, and summing across all observations (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 their Count data
- 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



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 SBN
- The most likely thresholds and correlations are estimated
- Genetic/Environmental variance components are estimated based on these correlations derived from MZ and DZ data

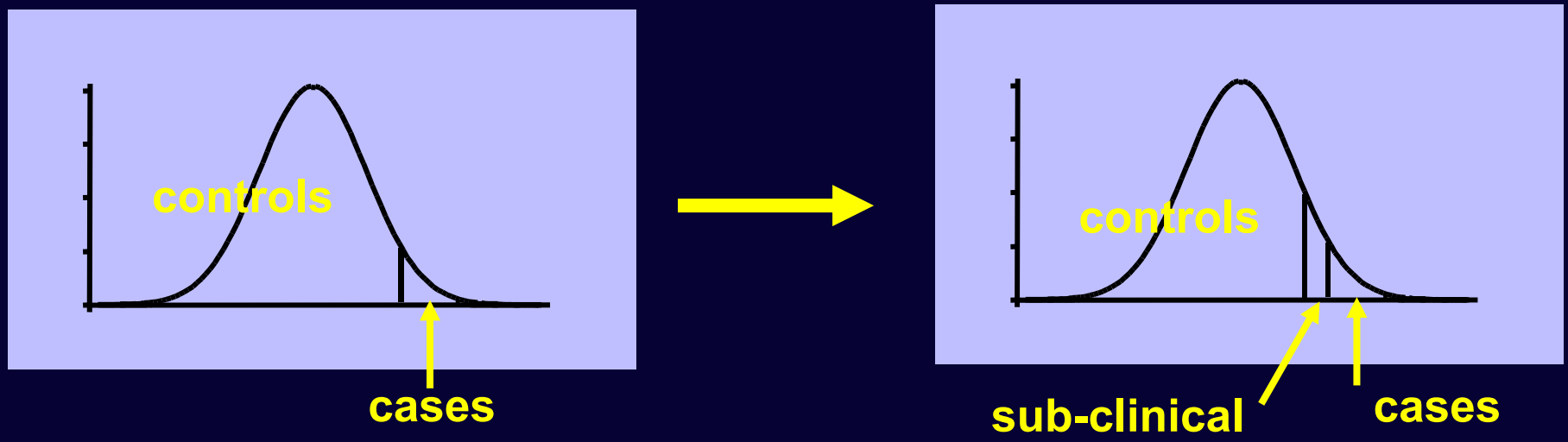
Power issues

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

Neale, Eaves & Kendler 1994

- Solutions:

1. Bigger samples
2. Use more categories



Practical

R Script: ThreshLiab.R
Data File: CASTage8.csv

Sample & Measures

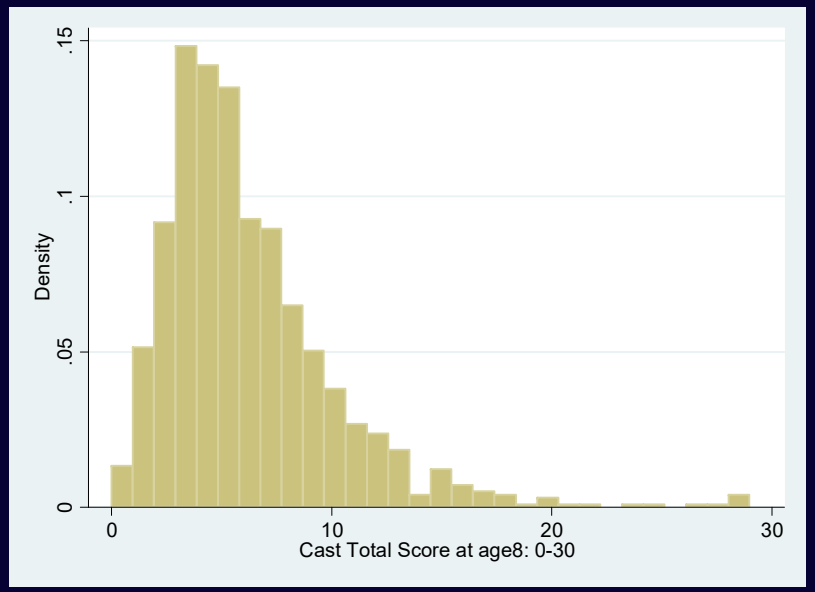
- Simulated data based on CAST data collected at age 8 in the TEDS sample
- Parent report of CAST: Childhood Autism Spectrum Test (Scott et al., 2002)
- Twin pairs: 501 MZ & 503 DZ males

Clinical Aspects of the CAST

The CAST score dichotomized at around 98% (i.e. scores of >15), is the clinical cut-off point for children at risk for Autism Spectrum Disorder

However, for the purpose of this exercise, we use 2 cut offs to create 3 categories:

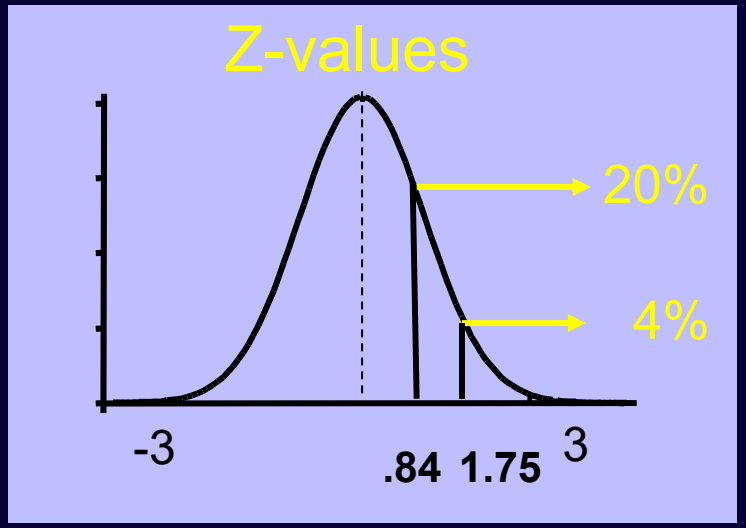
<9	unaffacted	(0)
9-15	sub-clinical	(1)
>15	ASD	(2)



Inspection of the data

CAST score categorized (0,1,2), the proportions:

CAST	Freq.	Percent
0	804	80.08
1	158	15.74
2	42	4.18
Total	1,004	100.00



Z-value Th1 = .84
Z-value Th2 = 1.75

CTs of the MZ and DZ group

MZ	0	1	2	Tot
0	385	23	6	414
1	28	37	4	69
2	3	3	12	18
Tot	416	63	22	501

table(mzData$Ocast1, mzDataF$Ocast2)

table(dzData$Ocast1, dzData$Ocast2)

DZ	0	1	2	Tot
0	334	37	19	390
1	56	32	1	89
2	12	2	10	24
Tot	402	81	30	503

R Script: ThreshLiab.R

```
Castdata <- read.table ('CASTage8.csv', header=T, sep=“,", na.strings=".",

selVars <- c(‘Ocast1' , ‘Ocast2')

# Declare variables to be ordered Factors for OpenMx
Castdata$Ocast1  <-mxFactor(Castdata$Ocast1, levels=c(0:2) )
Castdata$Ocast2  <-mxFactor(Castdata$Ocast2, levels=c(0:2) )

# Select Data for Analysis
mzData <- subset(Castdata, zyg==1, selVars)
dzData <- subset(Castdata, zyg==2, selVars)

# get CT for Ordinal variable
table(mzData$Ocast1, mzData$Ocast2)
table(dzData$Ocast1, dzData$Ocast2)
```

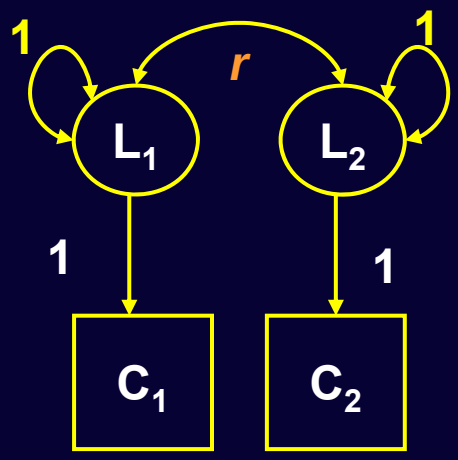
```
# 1) Specify Saturated Model (max number of parameters: 2 cor, 8 TH)
#     Matrices for expected Means (SND) & Tetrachoric correlations

meanL  <-mxMatrix( type="Zero", nrow=1, ncol=ntv, name="M" )
```

→ [0,0]

```
corMZ   <-mxMatrix(type="Stand", nrow=ntv, ncol=ntv, free=T, values=.8,
lbound=-.99, ubound=.99, name="expCorMZ")

CorDZ   <-mxMatrix(type="Stand", nrow=ntv, ncol=ntv, free=T, values=.8,
lbound=-.99, ubound=.99, name="expCorDZ")
```

	L1	L2
L1	1	*rMZ*
L2	*rMZ*	1

	L1	L2
L1	1	*rDZ*
L2	*rDZ*	1

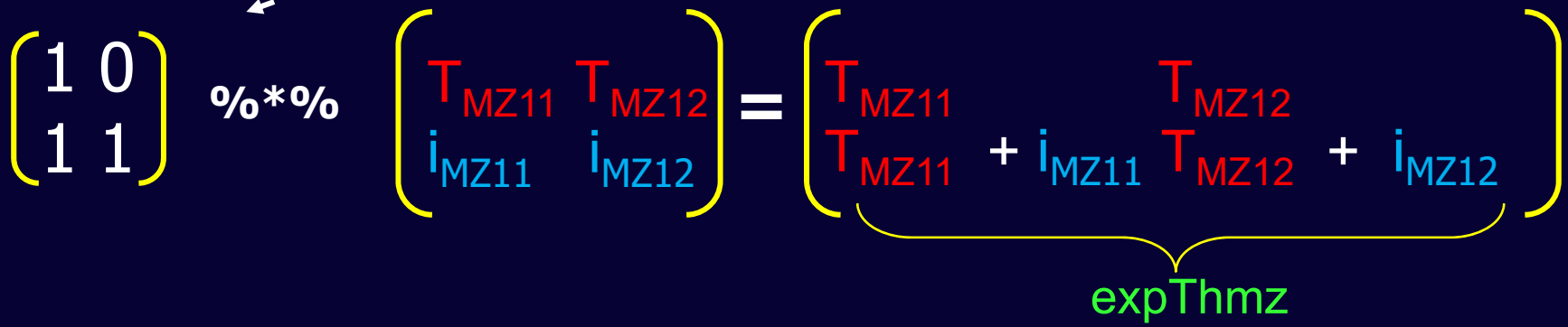


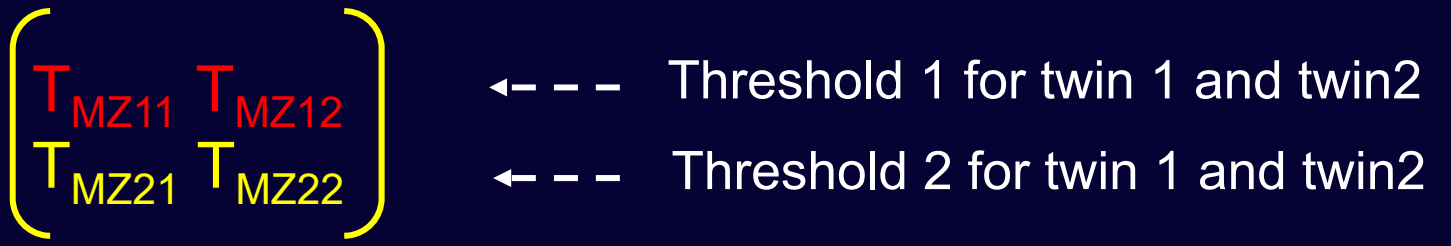
Matrices & Algebra for expected Thresholds

```
Tmz <-mxMatrix        (type="Full", nrow=nth, ncol=ntv, free=TRUE,
                      values=c(.8, 1,  .8, 1),
                      lbound=c(-3, .001, -3, .001 ),
                      ubound=(3),
                      labels=c("Tmz11","imz11", "Tmz12","imz12"),
                      name="ThMZ" )
```

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

Expected Thresholds: L %*% ThMZ





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

Start Values & Bounds

```
Tmz <-mxMatrix        (type="Full", nrow=nth, ncol=ntv, free=TRUE,
                      values=c(.8, 1,  .8, 1),
                      lbound=c(-3, .001, -3, .001 ),
                      ubound=(3),
                      labels=c("Tmz11","imz11", "Tmz12","imz12"),
                      name="ThMZ" )
```

$$\begin{pmatrix} T_{MZ11} & T_{MZ12} \\ i_{MZ11} & i_{MZ12} \end{pmatrix} = \begin{matrix} .8\ (-3 \text{ to } 3) & .8\ (-3 \text{ to } 3) \\ 1\ (.001 \text{ to } 3) & 1\ (.001 \text{ to } 3) \end{matrix}$$

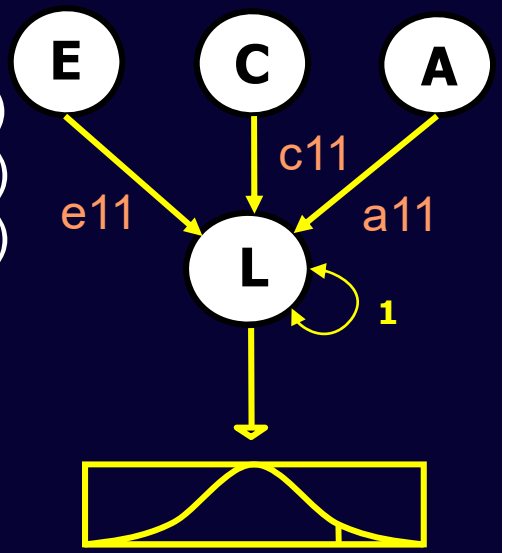
The positive bounds on the increments stop the thresholds going 'backwards', i.e. they preserve the ordering of the categories

Z-value Th1 = .84
Z-value Th2 = 1.75

```
# RUN SUBMODELS
# SubModel 1: Thresholds across Twins within zyg group are equal

Sub1Model    <- mxModel(SatModel, name="sub1")
Sub1Model    <- omxSetParameters( Sub1Model,
labels=c("Tmz11", "imz11", "Tmz12", "imz12"), newlabels=c("Tmz11", "imz11",
"Tmz11", "imz11"), …

# SubModel 3: Thresholds across Twins & zyg group are equal

Sub3Model    <- mxModel(Sub1Model, name="sub3")
Sub3Model    <- omxSetParameters( Sub3Model,
labels=c("Tdz11", "idz11", "Tdz12", "idz12"), newlabels=c("Tmz11", "imz11",
"Tmz11", "imz11"), …
```

omxSetParameters: function to modify the attributes of parameters in a model
Without having to re-specify the model

```
# ACE MODEL with one overall set of Thresholds
pathA    <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=.6,
label="a11", name="a" )
pathC    <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=.6,
label="c11", name="c" )
pathE    <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=.6,
label="e11", name="e" )

# Algebra for Matrices to hold A, C, and E Variance Components
covA     <- mxAlgebra( expression=a %*% t(a), name="A" )
covC     <- mxAlgebra( expression=c %*% t(c), name="C" )
covE     <- mxAlgebra( expression=e %*% t(e), name="E" )
covP     <- mxAlgebra( expression=A+C+E, name="V" )
```




```
# Constrain Total variance of the liability to 1
matUnv <-mxMatrix( type="Unit", nrow=nv, ncol=1,
name="Unv" )
varL     <-mxConstraint( expression=diag2vec(V)==Unv, name="VarL" )
```

A + C + E = 1

Practical

- Run first part of the script up to sub3Model
  - What are the conclusions about the thresholds, i.e. what is the best model?
  - What kind of Genetic model would you run on this data given the correlations?

- Run the ACE model and check the parameter estimates (with 95% CI)

MODEL	ep	-2LL	df	$\Delta\chi^2$(df)	P-val
1 All TH free	10	2202.7	1998	-	-
2 Sub1: TH tw1=tw2 in MZ	8	2203.8	2000	1.01 (2)	.61 ns
3 Sub2: TH tw1=tw2 in DZ	8	2206.0	2000	3.24 (2)	.20 ns
4 Sub3: One overall TH	4	2211.1	2004	8.40 (6)	.21 ns

1 Thresh/Inc:	MZ tw1 = .94, .84	MZ tw2 = .96, .73
	DZ tw1 = .75, .91	DZ tw2 = .84, .71
2 Thresh/Inc:	MZ = .95, .78	
	DZ tw1 = .75, .91	DZ tw2 = .84, .71
3 Thresh/Inc:	MZ tw1 = .94, .84	MZ tw2 = .96, .73
	DZ = .80, .81	
4 Thresh/Inc:	.86, .79	

The Twin correlations for model 4 are:

r_{MZM} = 0.82 (.74 - .87) r_{DZM} = 0.44 (.30 - .56)

ACE Estimates for the ordinalized CAST score in Boys at age 8

	h^2	c^2	e^2
ACE	.76 (.48/.87)	.06 (0/.31)	.18 (.13/.26)

	Name	ep	-2LL	df	AIC
Model 1 :	ACE	5	2211.14	2004	-1796.86

Age Effects on Thresholds

- The effect of covariates like Age can be modelled in the Threshold model, similarly to the means model
- An example script is added to the folder in which Age is incorporated and its effects modelled in the thresholds (Age Regression on TH.R)

$$\begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} + \begin{pmatrix} \text{BageTH*Age1} & \text{BageTH*Age2} \\ \text{BageTH*Age1} & \text{BageTH*Age2} \end{pmatrix}$$