Categorical Data

Frühling Rijsdijk & Kate Morley

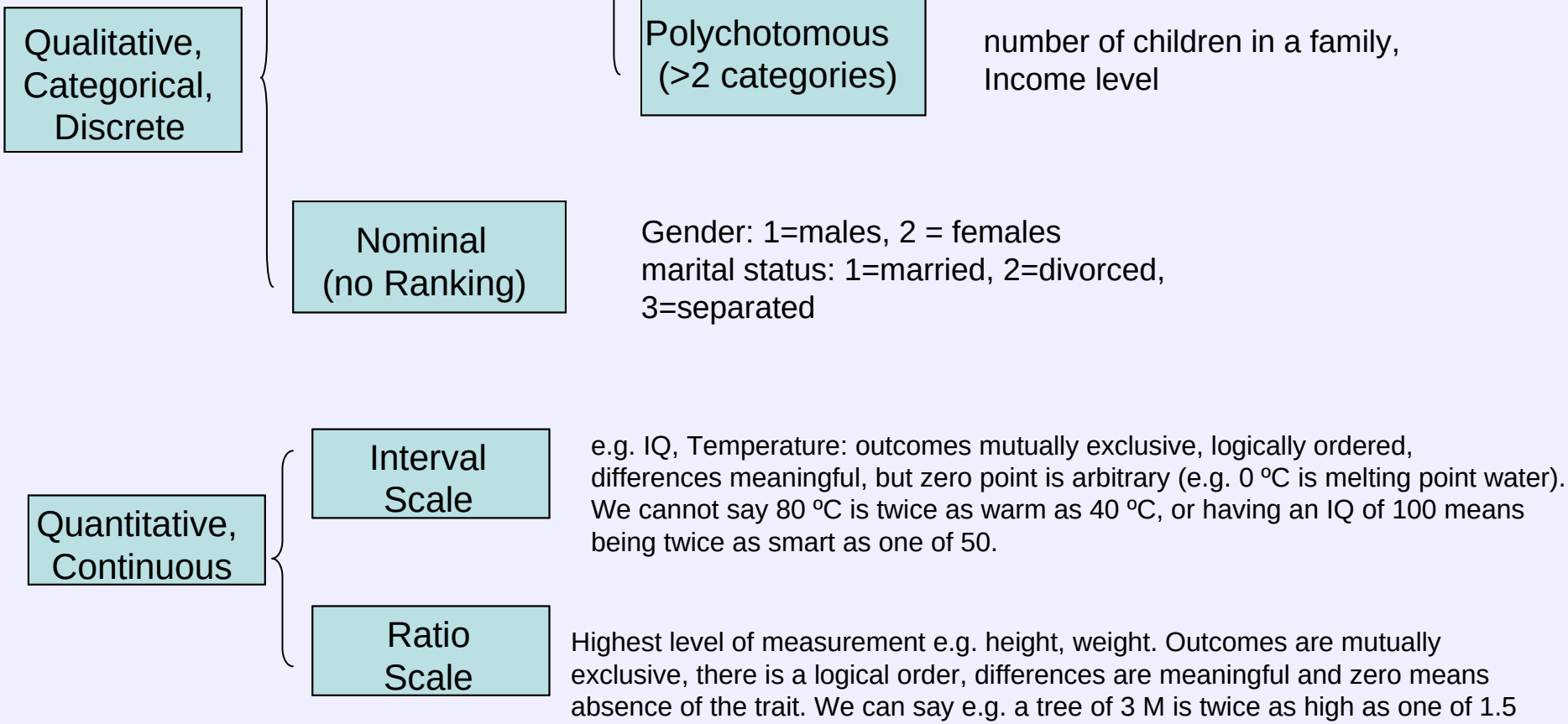
Twin Workshop, Boulder

Tuesday March 7, 2006

Aims

- Introduce Categorical Data
- Define liability and describe assumptions of the liability model
- Show how heritability of liability can be estimated from categorical twin data
- Practical exercises

Measurement Scales of Outcome Variables



Ordinal data

Measuring instrument is able to only discriminate between two or a few ordered categories e.g. absence or presence of a disease. Data take the form of counts, i.e. the number of individuals within each category:

Of 100 individuals:

90 ‘no’
10 ‘yes’

	‘no’	**‘yes’**
‘no’	**55**	**19**
‘yes’	**18**	**8**

Univariate Normal Distribution of Liability

Assumptions:

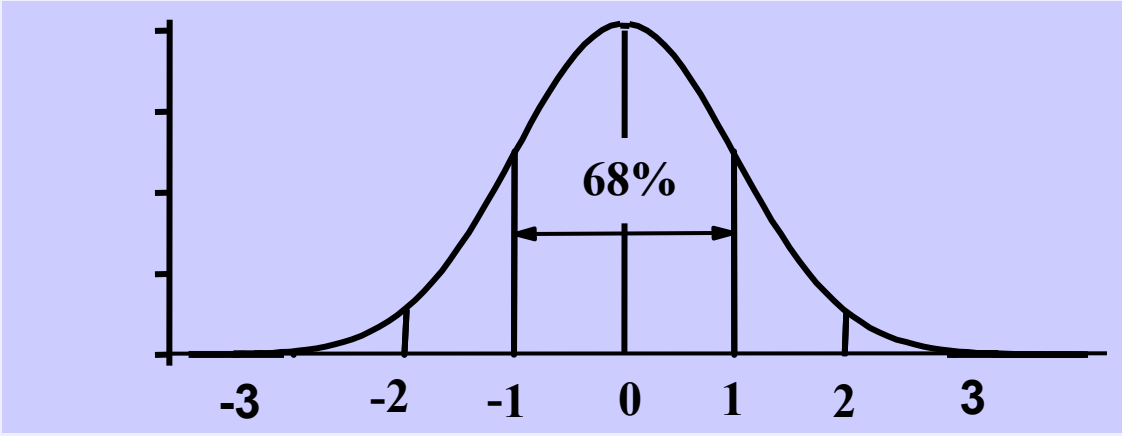
(1) Underlying *normal* distribution of liability

(2) The liability distribution has 1 or more thresholds (cut-offs)

The standard Normal distribution

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

- mean (μ) = 0 and SD (σ) = 1
- z-values are the number of SD away from the mean
- area under curve translates directly to probabilities > Normal Probability Density function (Φ)

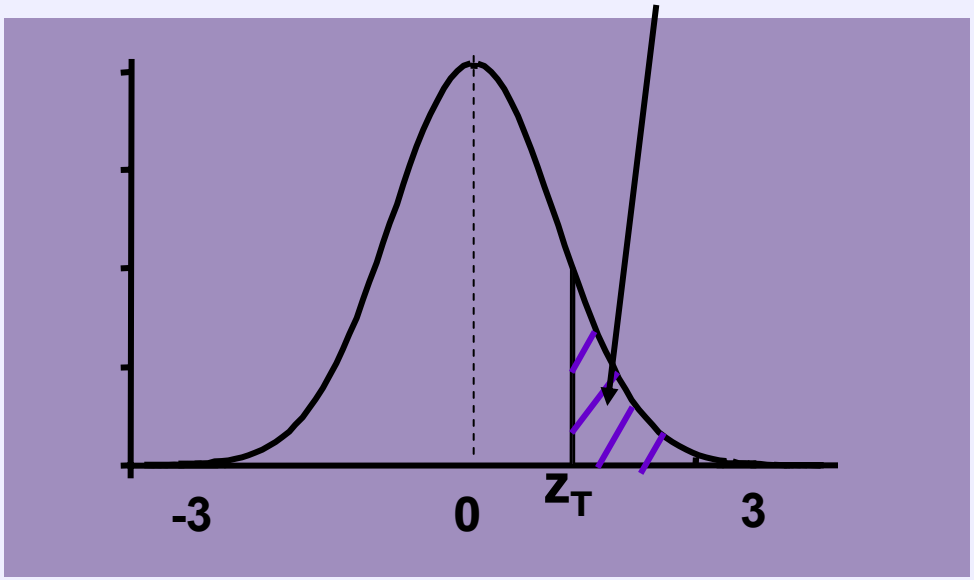


Standard Normal Cumulative Probability in right-hand tail

(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($z \geq z_T$)

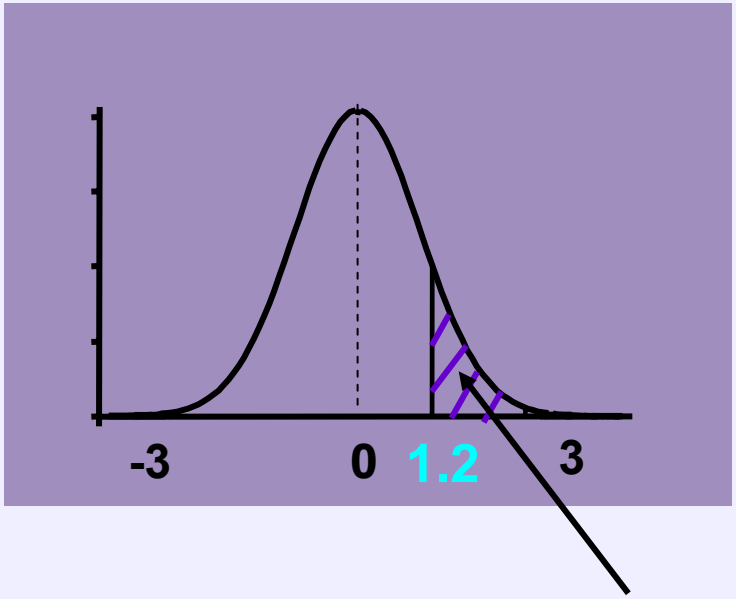


$$\int_{z_T}^{\infty} \Phi(L_1; \mu = 0, \sigma^2 = 1)\, dL_1$$

Example: From counts find z-value in Table

For one variable it is possible to find a **z-value** (threshold) on the SND, so that the proportion exactly matches the observed proportion of the sample e.g. if from a sample of **1000** individuals, **120** have met a criteria for a disorder (**12%**): the z-value is **1.2**

Z_0	Area	
.6	.27	27%
.8	.21	21%
1	.16	16%
1.2	.12	12%
1.4	.08	8%
1.6	.055	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%



Two categorical traits:
Data from twins

In an unselected sample of twins > Contingency Table with 4 observed cells:

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

Twin2 \ Twin1	0	1
0	545 (.76)	75 (.11)
1	56 (.08)	40 (.05)

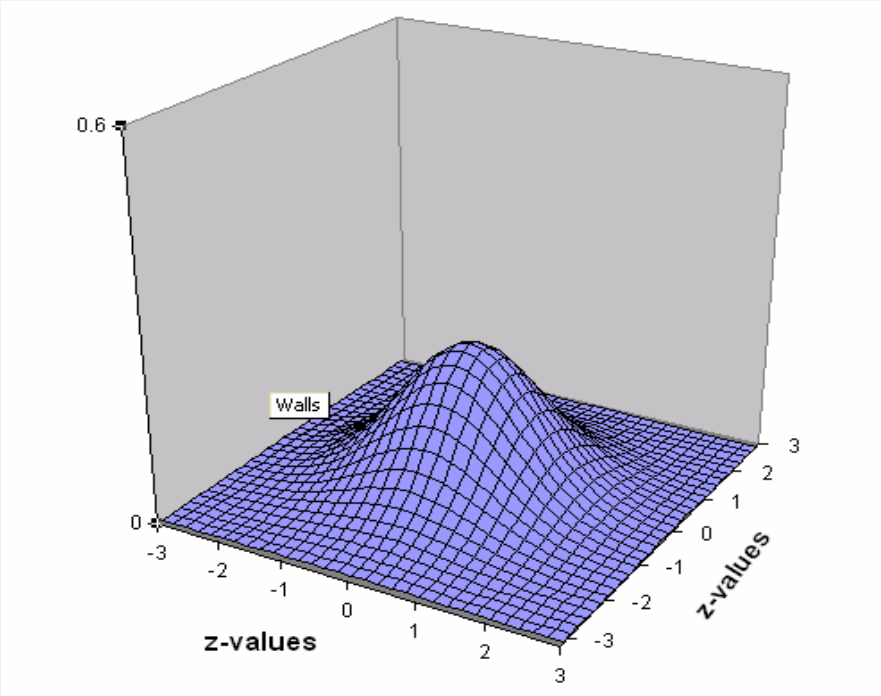
0 = unaffected
1 = affected

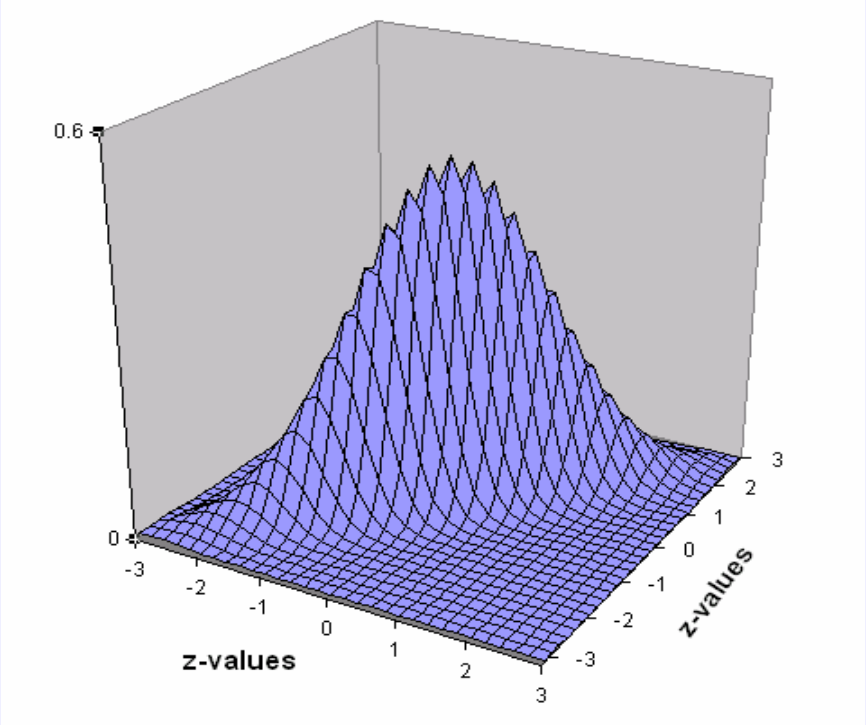
Joint Liability Model for twin pairs

- 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 unknown.

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

Bivariate Normal

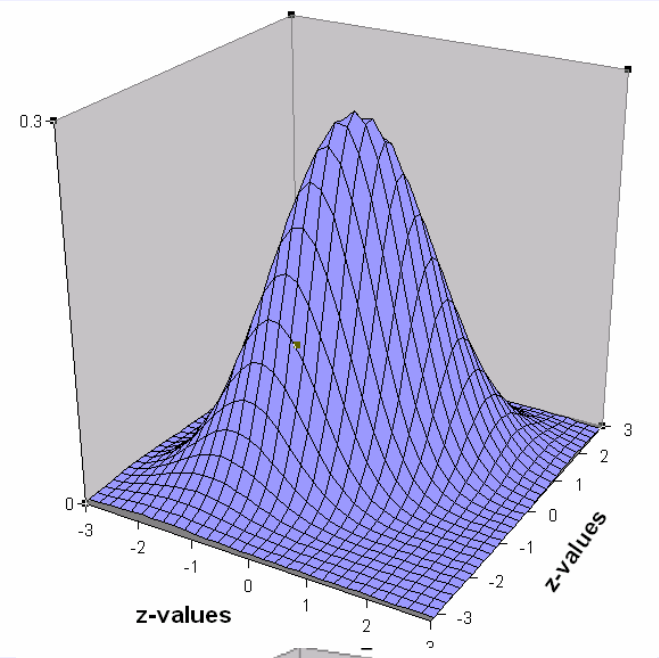


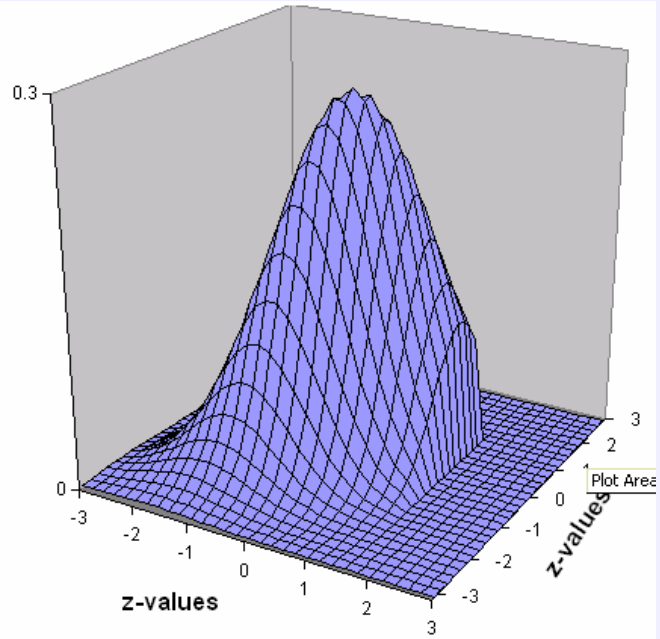


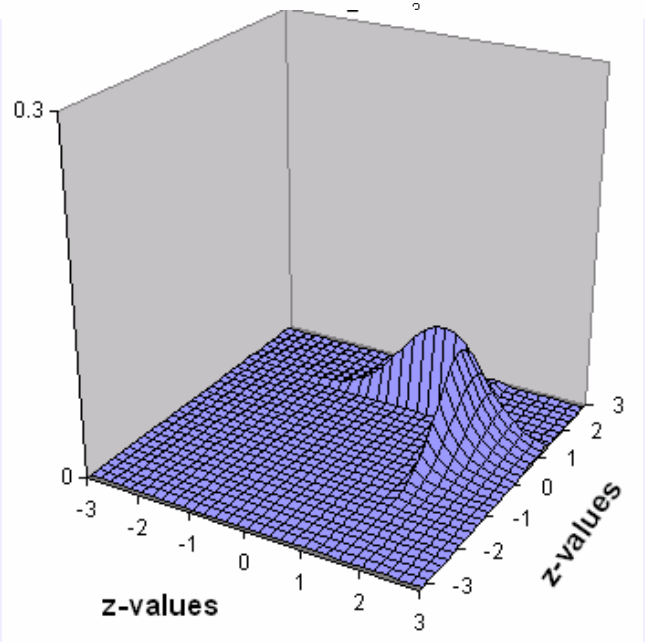
***r* = .00**

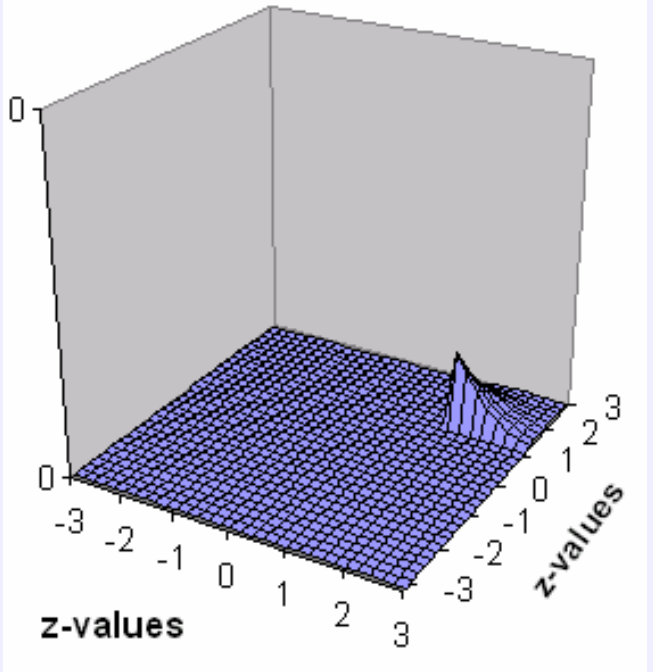
***r* = .90**

Bivariate Normal (R=0.6) partitioned at threshold 1.4 (z-value) on both liabilities









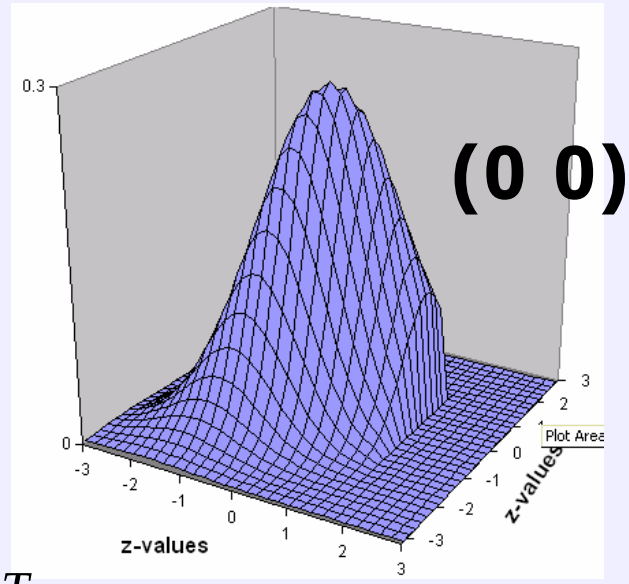
How are expected proportions calculated?

By **numerical integration** of the bivariate normal over two dimensions: the liabilities for twin1 and twin2

e.g. the probability that both twins are affected :

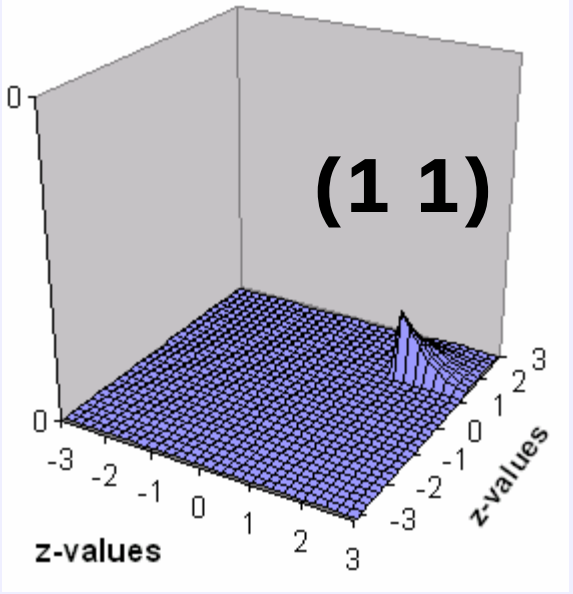
$$\int_{T_1}^{\infty}\int_{T_2}^{\infty} \Phi(L_1, L_2; \mathbf{0}, \boldsymbol{\Sigma})\, dL_1 dL_2$$

Φ is the bivariate normal probability density function, L_1 and L_2 are the liabilities of twin1 and twin2, with means 0, and Σ is the correlation matrix of the two liabilities
T_1 is threshold (z-value) on L_1, T_2 is threshold (z-value) on L_2



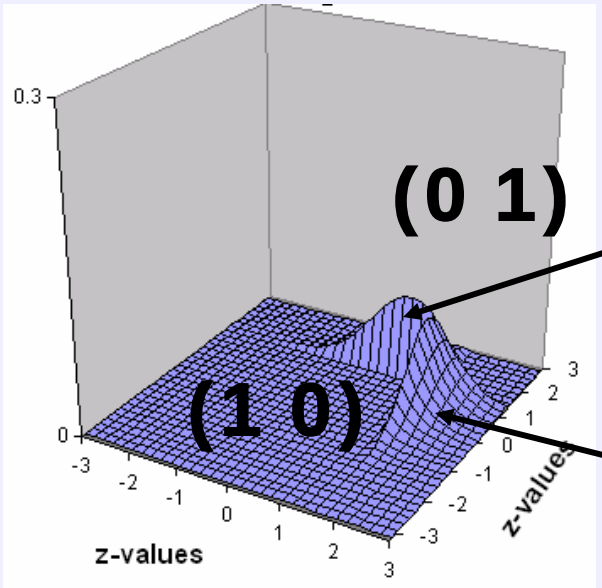
(0 0)

$$\int_{-\infty}^{T_1}\int_{-\infty}^{T_2}\Phi(L_1,L_2;\mathbf{0},\boldsymbol{\Sigma})\,dL_1dL_2$$



(1 1)

$$\int_{T_1}^{\infty}\int_{T_2}^{\infty}\Phi(L_1,L_2;\mathbf{0},\boldsymbol{\Sigma})\,dL_1dL_2$$



(0 1)

(1 0)

$$\int_{-\infty}^{T_1}\int_{T_2}^{\infty}\Phi(L_1,L_2;\mathbf{0},\boldsymbol{\Sigma})\,dL_1dL_2$$

$$\int_{T_1}^{\infty}\int_{-\infty}^{T_2}\Phi(L_1,L_2;\mathbf{0},\boldsymbol{\Sigma})\,dL_1dL_2$$

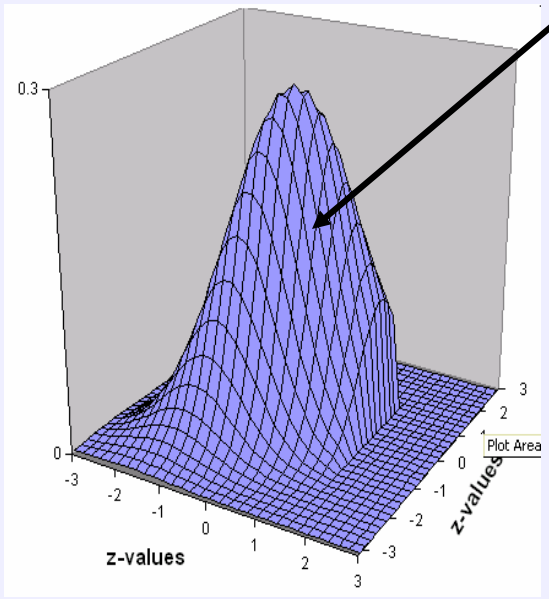
How is numerical integration performed?

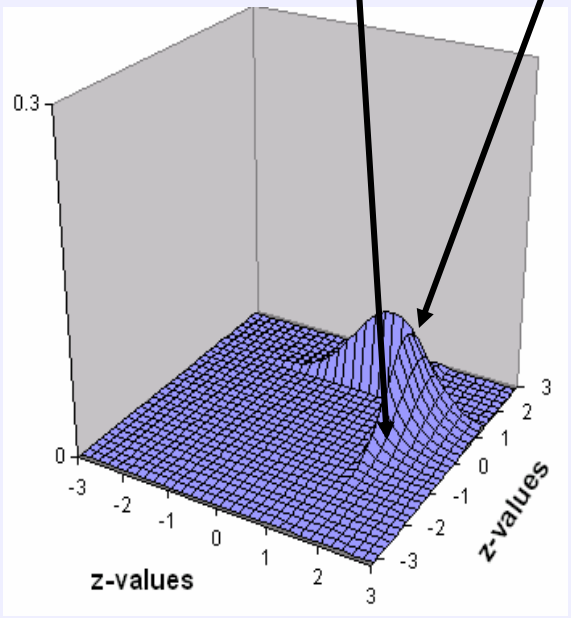
There are programmed mathematical subroutines that can do these calculations

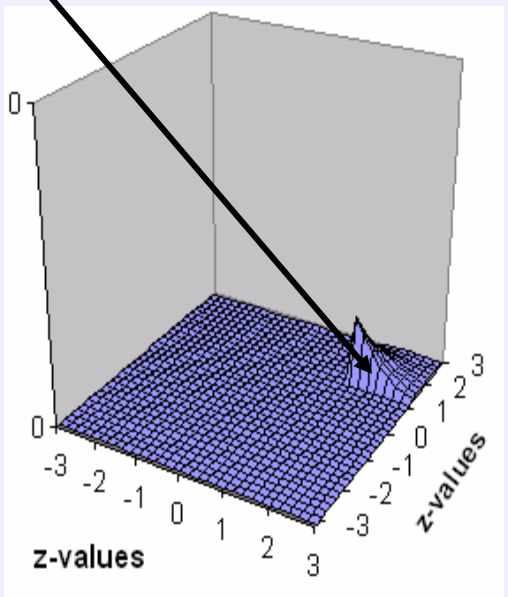
Mx uses one of them

Expected Proportions of the BN, for R=0.6, Th1=1.4, Th2=1.4

Liab 1 \ Liab 2	0	1
0	.87	.05
1	.05	.03





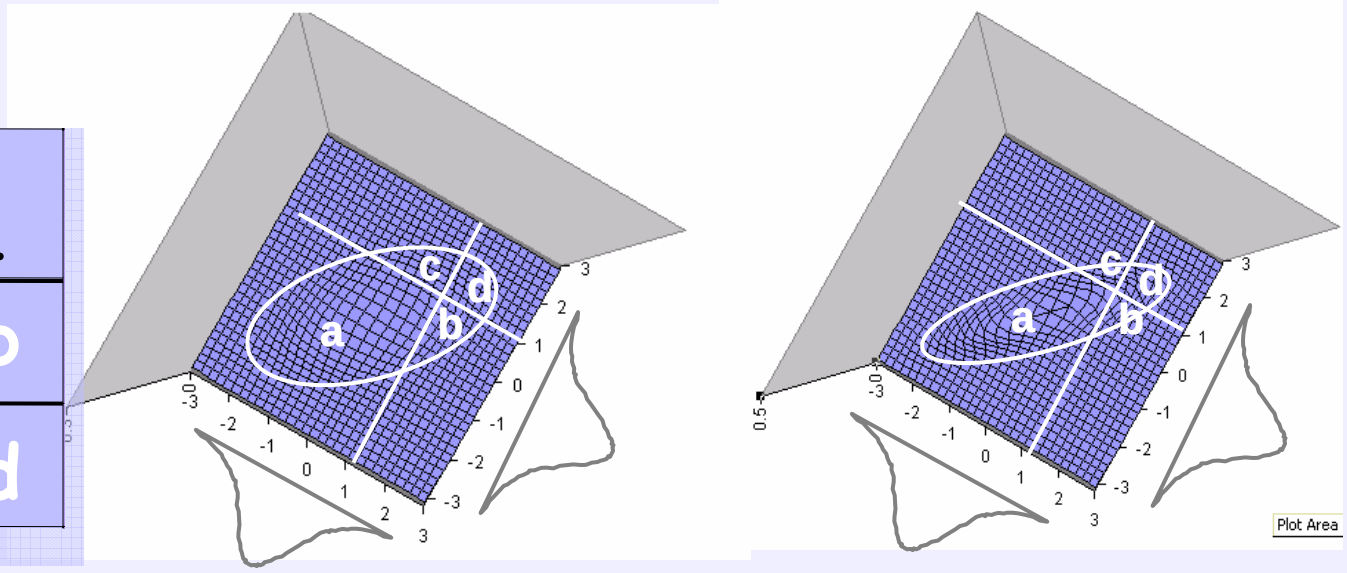


How can we estimate correlations from CT?

The correlation (shape) of the BN and the two thresholds determine the relative proportions of observations in the 4 cells of the CT.

Conversely, the sample proportions in the 4 cells can be used to estimate the correlation and the thresholds.

Twin1 \ Twin2	0	1
0	a	b
1	c	d

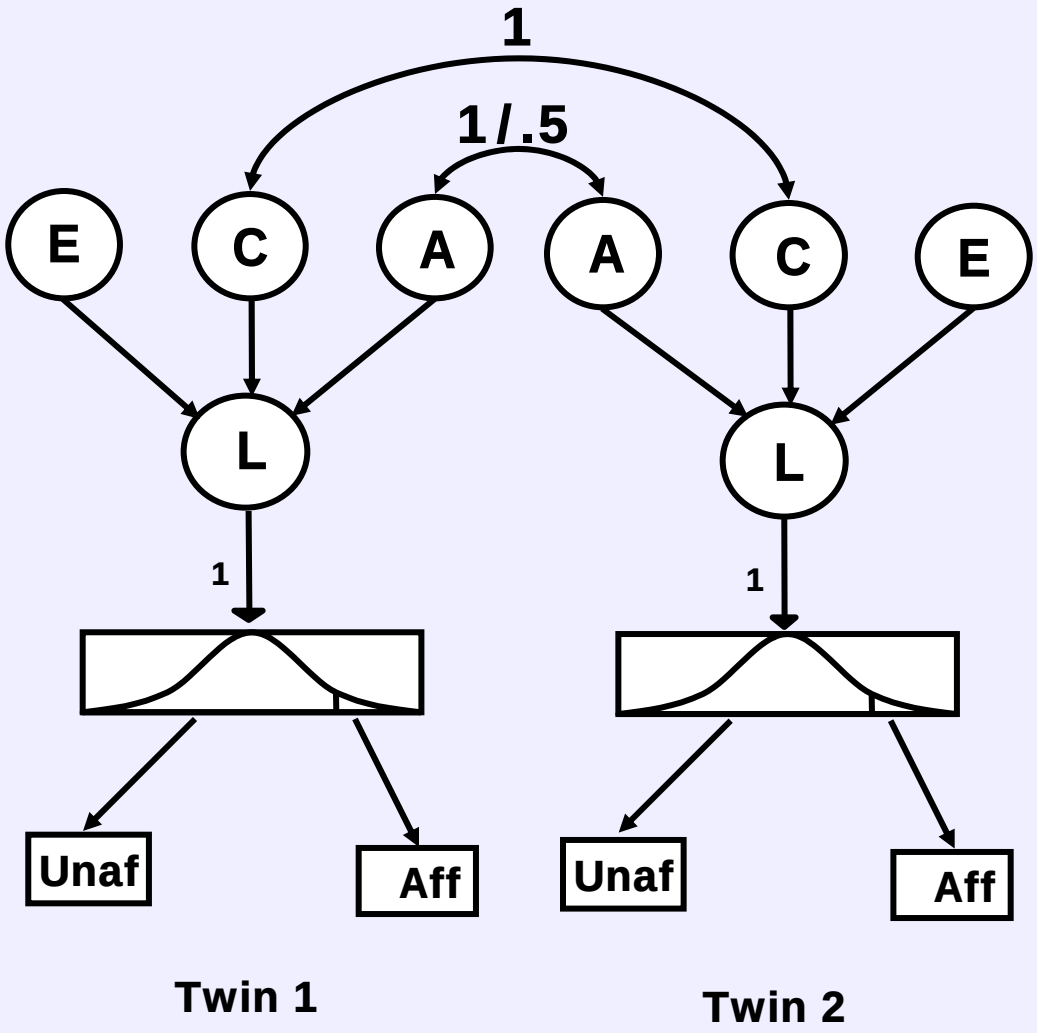


Summary

It is possible to estimate a correlation between categorical traits from simple counts (CT), because of the assumptions we make about their joint distributions: The Bivariate Normal

The relative sample proportions in the 4 cells are translated to proportions under the BN so that the most likely correlation and the thresholds are derived

ACE Liability Model



How can we fit ordinal data in Mx?

Summary statistics: CT

Mx has a built-in fit function for the maximum-likelihood analysis of 2-way Contingency Tables

>analyses limited to only two variables

Raw data analyses

- multivariate
- handles missing data
- moderator variables

ML of RAW Ordinal data

Is the sum of the likelihood of all observations. The likelihood of an observation is the expected proportion in the corresponding cell of the MN.

The sum of the log-likelihoods of all observations is a value that (like for continuous data) is not very interpretable, unless we compare it with the LL of other models or a saturated model to get a chi-square index.

Raw Ordinal Data

Zyg	ordinal respons1	ordinal respons2
1	0	0
1	0	0
1	0	1
2	1	0
2	0	0
1	1	1
2	.	1
2	0	.
2	0	1

NOTE: smallest category should always be 0 !!

SORT !

We can speed up computation time considerably when the data is sorted since if case i+1 = case i, then likelihood is NOT recalculated.

In e.g. the bivariate, 2 category case, there are only 4 possible vectors of observations :

1 1, 0 1, 1 0, 00 and, therefore, only 4 integrals for Mx to calculate if the data file is sorted.

Practical

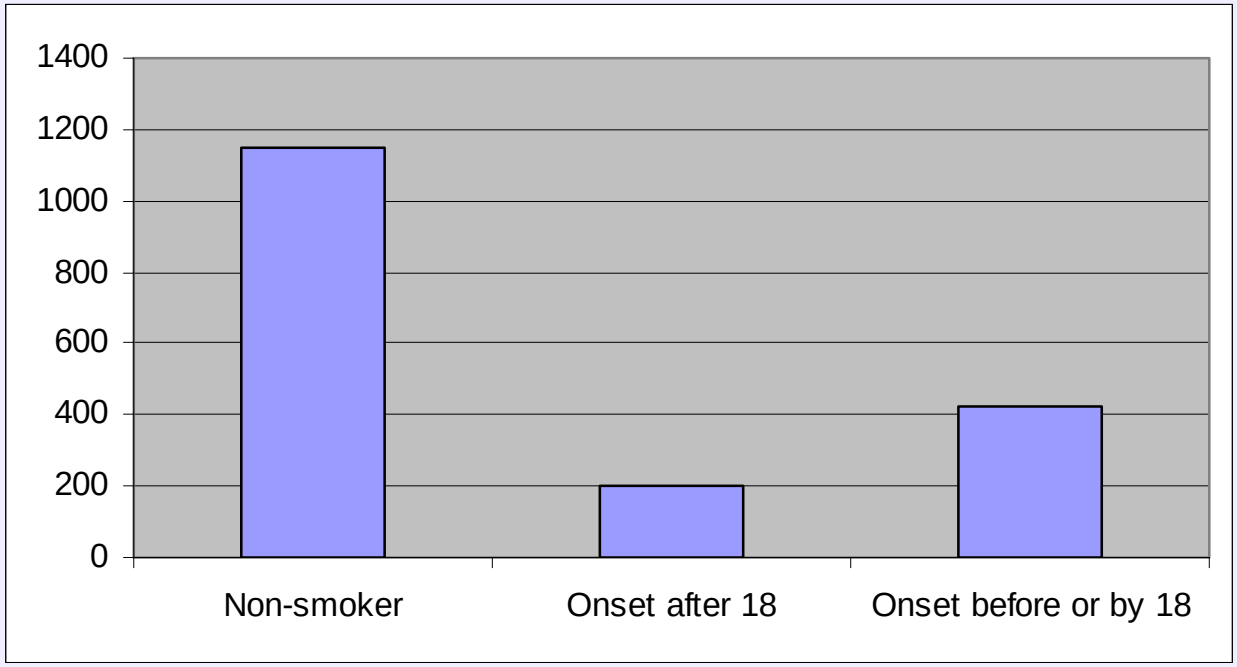
Sample and Measures

- Australian Twin Registry data (QIMR)
- Self-report questionnaire
  - Non-smoker, ex-smoker, current smoker
  - Age of smoking onset
- Large sample of adult twins

  + family members
  - Today using MZMs (785 pairs)

    and DZMs (536 pairs)

- Variable: age at smoking onset, including non-smokers
- Ordered as:
  - Non-smokers / late onset / early onset



Practical Exercise

Analysis of age of onset data
- **Estimate thresholds**
- **Estimate correlations**
- **Fit univariate model**

Observed counts from ATR data:

MZM

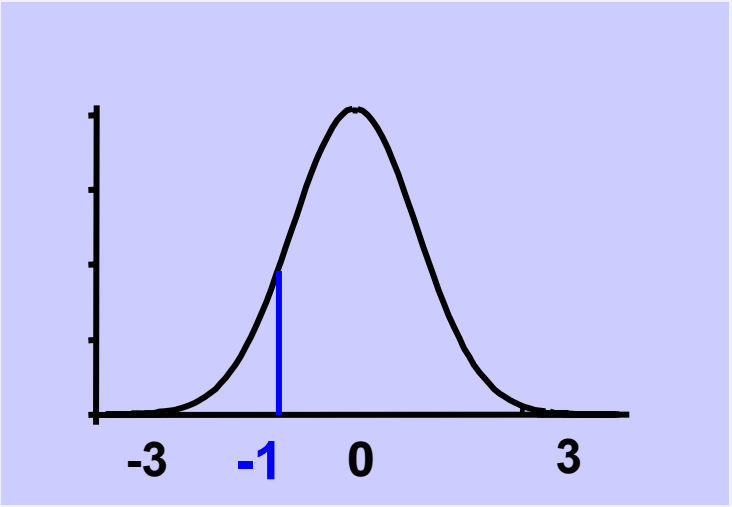
	0	1	2
0	368	24	46
1	26	15	21
2	54	22	209

DZM

	0	1	2
0	203	22	63
1	17	5	16
2	65	12	133

Threshold Specification in Mx

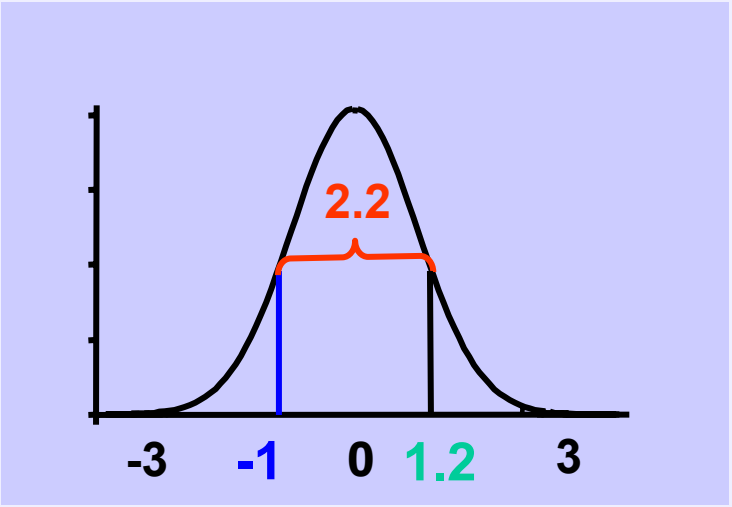
2 Categories
Matrix T: 1 x 2
T(1,1) T(1,2) **threshold 1 for twin1 & twin2**



Threshold Model T /

Threshold Specification in Mx

3 Categories
Matrix T: 2 x 2
T(1,1) T(1,2) threshold 1 for twin1 & twin2
T(2,1) T(2,2) increment



Threshold Model L*T /

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ t_{11} + t_{21} & t_{12} + t_{22} \end{bmatrix}$$

polycor_smk.mx

```
#define nvarx2 2
#define nthresh 2
#ngroups 2

G1: Data and model for MZM correlation
DAta NInput_vars=3
Missing=.
Ordinal File=smk_prac.ord
Labels
  zyg ageon_t1 ageon_t2
SELECT IF zyg = 2
SELECT ageon_t1 ageon_t2 /

Begin Matrices;
R STAN nvarx2 nvarx2 FREE
T FULL nthresh nvarx2 FREE
L Lower nthresh nthresh
End matrices;

Value 1 L 1 1 to L nthresh nthresh
```

polycor_smk.mx

```
#define nvarx2 2                    ! Number of variables x number of twins
#define nthresh 2                   ! Number of thresholds=num of cat-1
#ngroups 2

G1: Data and model for MZM correlation
DAta NInput_vars=3
Missing=.
Ordinal File=smk_prac.ord           ! Ordinal data file
Labels
  zyg ageon_t1 ageon_t2
SELECT IF zyg = 2
SELECT ageon_t1 ageon_t2 /

Begin Matrices;
R STAN nvarx2 nvarx2 FREE
T FULL nthresh nvarx2 FREE
L Lower nthresh nthresh
End matrices;

Value 1 L 1 1 to L nthresh nthresh
```

polycor_smk.mx

```
#define nvarx2 2                    ! Number of variables per pair
#define nthresh 2                   ! Number of thresholds=num of cat-1
#ngroups 2

G1: Data and model for MZM correlation
DAta NInput_vars=3
Missing=.
Ordinal File=smk_prac.ord           ! Ordinal data file
Labels
  zyg ageon_t1 ageon_t2
SELECT IF zyg = 2
SELECT ageon_t1 ageon_t2 /

Begin Matrices;
R STAN nvarx2 nvarx2 FREE           ! Correlation matrix
T FULL nthresh nvarx2 FREE
L Lower nthresh nthresh
End matrices;

Value 1 L 1 1 to L nthresh nthresh
```

polycor_smk.mx

```
#define nvarx2 2                    ! Number of variables per pair
#define nthresh 2                   ! Number of thresholds=num of cat-1
#ngroups 2

G1: Data and model for MZM correlation
DAta NInput_vars=3
Missing=.
Ordinal File=smk_prac.ord           ! Ordinal data file
Labels
  zyg ageon_t1 ageon_t2
SELECT IF zyg = 2
SELECT ageon_t1 ageon_t2 /

Begin Matrices;
R STAN nvarx2 nvarx2 FREE           ! Correlation matrix
T FULL nthresh nvarx2 FREE          ! thresh tw1, thresh tw2
L Lower nthresh nthresh             ! Sums threshold displacements
End matrices;

Value 1 L 1 1 to L nthresh nthresh          ! initialize L
```

```
COV                R /

Thresholds         L*T /

Bound 0.01 1 T 1 1 T 1 2
Bound 0.1 5 T 2 1 T 2 2
Start 0.2 T 1 1 T 1 2
Start 0.2 T 2 1 T 2 2
Start .6 R 2 1
Option RS
Option func=1.E-10
END
```

```
COV              R /          ! Predicted Correlation matrix for MZ pairs

Thresholds       L*T /        ! Threshold model, to ensure t1>t2>t3 etc.....

Bound 0.01 1 T 1 1 T 1 2
Bound 0.1 5 T 2 1 T 2 2
Start 0.2 T 1 1 T 1 2
Start 0.2 T 2 1 T 2 2
Start .6 R 2 1
Option RS
Option func=1.E-10
END
```

```
COV             R /          ! Predicted Correlation matrix for MZ pairs

Thresholds      L*T /        ! Threshold model, to ensure t1>t2>t3 etc.....

Bound 0.01 1 T 1 1 T 1 2
Bound 0.1 5 T 2 1 T 2 2      ! Ensures positive threshold displacement
Start 0.2 T 1 1 T 1 2        ! Starting values for the 1st thresholds
Start 0.2 T 2 1 T 2 2        ! Starting values for the 2nd thresholds
Start .6 R 2 1               ! Starting value for the correlation
Option RS
Option func=1.E-10           !function precision is less than usual
END
```

```
! Test equality of thresholds between Tw1 and Tw2
EQ T 1 1 1 T 1 1 2          !constrain TH1 to be equal across Tw1 and Tw2 MZM
EQ T 1 2 1 T 1 2 2          !constrain TH2 to be equal across Tw1 and Tw2 MZM
EQ T 2 1 1 T 2 1 2          !constrain TH1 to be equal across Tw1 and Tw2 DZM
EQ T 2 2 1 T 2 2 2          !constrain TH2 to be equal across Tw1 and Tw2 DZM
End

Get cor.mxs

! Test equality of thresholds between MZM & DZM
EQ T 1 1 1 T 1 1 2 T 2 1 1 T 2 1 2   !constrain TH1 to be equal across all Males
EQ T 1 2 1 T 1 2 2 T 2 2 1 T 2 2 2   !constrain TH2 to be equal across all Males
End
```

Exercise I

- Fit saturated model
  - Estimates of thresholds
  - Estimates of polychoric correlations
- Test equality of thresholds
  - Examine differences in threshold and correlation estimates for saturated model and sub-models
- Examine correlations
  - What model should we fit?

Raw ORD File: **smk_prac.dat**
Script: **polychor_smk.mx**
Location: **kate\Ordinal_Practical**

Estimates: smoking age-at-onset

-2LL	df				Twin 1	Twin 2		Twin 1	Twin 2
Saturated									
5128.185	3055		Th1	MZ	0.09	0.12	DZ	0.03	0.05
			Th2		0.31	0.33		0.24	0.26
			Cor		1			1	
					0.81	1		0.55	1

Estimates: smoking age-at-onset

ΔX^2	Δ df	P			Twin 1	Twin 2		Twin 1	Twin 2
Sub-model 1									
			Th1	MZ			DZ		
			Th2						
			Cor						
Sub-model 2									
			Th1	MZ			DZ		
			Th2						
			Cor						

Estimates: smoking age-at-onset

ΔX^2	Δ df	P			Twin 1	Twin 2		Twin 1	Twin 2
Sub-model 1									
0.77	4	0.94	Th1	MZ	0.10	0.10	DZ	0.04	0.04
			Th2		0.32	0.32		0.25	0.25
			Cor		1			1	
					0.81	1		0.55	1
Sub-model 2									
2.44	6	0.88	Th1	MZ	0.07	0.07	DZ	0.07	0.07
			Th2		0.29	0.29		0.29	0.29
			Cor		1			1	
					0.81	1		0.55	1

ACEcat_smk.mx

```
#define nvar 1                          ! number of variables per twin
#define nvarx2 2                        ! number of variables per pair
#define nthresh 1                       ! number of thresholds=num of cat-1
#ngroups 4                              ! number of groups in script

G1: Parameters for the Genetic model
Calculation
Begin Matrices;
X Low nvar nvar FREE                    ! Additive genetic path coefficient
Y Low nvar nvar FREE                    ! Common environmental path coefficient
Z Low nvar nvar FREE                    ! Unique environmental path coefficient
End matrices;
Begin Algebra;
A=X*X' ;                                !Additive genetic variance (path X squared)
C=Y*Y' ;                                !Common Environm variance (path Y squared)
E=Z*Z' ;                                !Unique Environm variance (path Z squared)
End Algebra;
start .6 X 1 1 Y 1 1 Z 1 1              !starting value for X, Y, Z
Interval @95 A 1 1 C 1 1 E 1 1          !requests the 95%CI for h2, c2, e2
End
```

```
G2: Data and model for MZ pairs
DAta NInput_vars=3
Missing=.
Ordinal File=prac_smk.ord
Labels
  zyg ageon_t1 ageon_t2
SELECT IF zyg = 2
SELECT ageon_t1 ageon_t2 /
Matrices = group 1
T FULL nthresh nvarx2 FREE      ! Thresh tw1, thresh tw2
L Lower nthresh nthresh

COV                             ! Predicted covariance matrix for MZ pairs
 ( A + C + E | A + C _
     A + C     | A + C + E ) /
Thresholds  L*T /               !Threshold model

Bound 0.01 1 T 1 1 T 1 2        ! Ensures positive threshold displacement
Bound 0.1 5 T 2 1 T 2 2
Start 0.1 T 1 1 T 1 2           ! Starting values for the 1st thresholds
Start 0.2 T 1 1 T 1 2           ! Starting values for the 2nd thresholds
Option rs
End
```

```
G3: Data and model for DZ pairs
DAta NInput_vars=4
Missing=.
Ordinal File=prac_smk.ord
Labels
  zyg ageon_t1 ageon_t2
SELECT IF zyg = 4
SELECT ageon_t1 ageon_t2 /
Matrices = group 1
T FULL nthresh nvarx2 FREE      ! Thresh tw1, thresh tw2
L Lower nthresh nthresh
H FULL 1 1                      ! .5

COVARIANCE                      ! Predicted covariance matrix for DZ pairs
 ( A + C + E | H@A + C _
   H@A + C   | A + C + E ) /
Thresholds  L*T /               !Threshold model

Bound 0.1 1 T 1 1 T 1 2         ! Ensures positive threshold displacement
Bound 0.1 5 T 2 1 T 2 2
Start 0.1 T 1 1 T 1 2           ! Starting values for the 1st thresholds
Start 0.2 T 1 1 T 1 2           ! Starting values for the 2nd thresholds
Option rs
End
```

G4: CONSTRAIN VARIANCES OF OBSERVED VARIABLES TO 1

```
CONSTRAINT
Matrices = Group 1
I UNIT 1 1
CO A+C+E= I  /                    !constrains the total variance to equal 1
Option func=1.E-10
End
```

Constraint groups and degrees of freedom

As the total variance is constrained to unity, we can estimate one VC from the other two, giving us one less independent parameter:

A + C + E = 1 therefore E = 1 - A - C

So each constraint group adds a degree of freedom to the model.

Exercise II

- Fit ACE model
  - What does the threshold model look like?
  - Change it to reflect the findings from exercise I

Raw ORD File: smk_prac.dat
Script: ACEcat_smk.mx
Location: kate\Ordinal_Practical