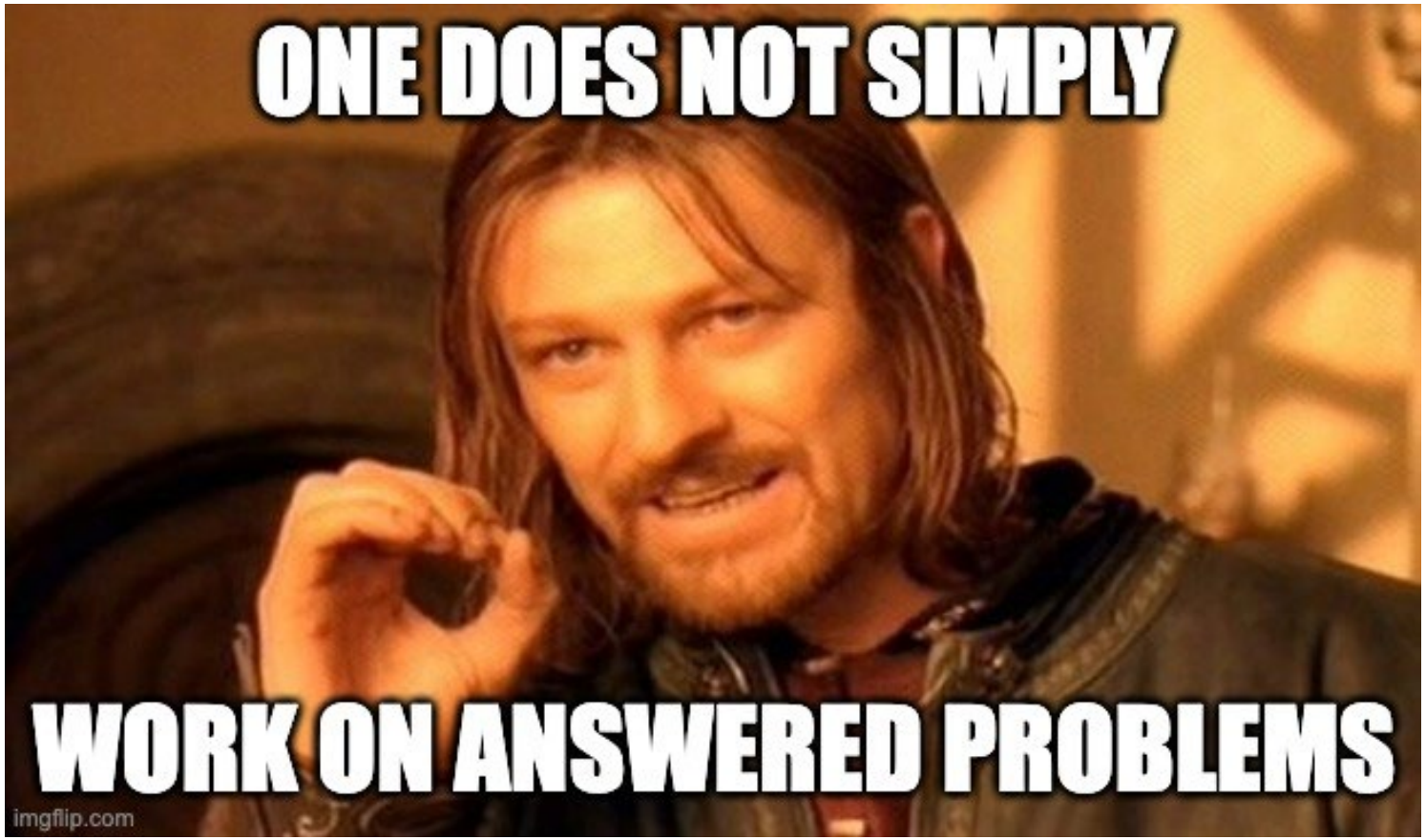
Model building, assumptions and the practice of science

Benjamin Neale, Ph.D.

International Statistical Genetics Workshop

March 4th 2024

All science starts with fantasy



An example model

- Consider a genotype, G
- What might we want to know about this genotype?

An example model

- Consider a genotype, G
- What might we want to know about this genotype?
- Number of distinct alleles?
- Frequency?
- Effect size?

An example model

- Consider a trait, T
- What might we want to know about this trait?

An example model

- Consider a trait, T
- What might we want to know about this trait?
- Mean?
- Distribution?
- Sources of variation?

An example model

- Consider our G and T (cheers)
- Let us say that T = $\boldsymbol{\beta}$*G + E

An example model

- Consider our G and T (cheers)
- Let us say that T = $\boldsymbol{\beta}$*G + E
- How do we figure out the heritability of T?

An example model

- Consider our G and T (cheers)
- Let us say that T = $\boldsymbol{\beta}$*G + E
- How do we figure out the heritability of T?
- Var(T) = Var($\boldsymbol{\beta}$ *G + E)
- Recall Var(X + Y) = Var(X) + Var(Y) + 2*Cov(X,Y)
- So Var(T) = Var($\boldsymbol{\beta}$ *G) + Var(E) + 2*Cov($\boldsymbol{\beta}$ *G,E)

An example model

- Consider our G and T (cheers)
- Let us say that T = $\boldsymbol{\beta}$*G + E
- How do we figure out the heritability of T?
- Var(T) = Var($\boldsymbol{\beta}$ *G + E)
- Recall Var(X + Y) = Var(X) + Var(Y) + 2*Cov(X,Y)
- So Var(T) = Var($\boldsymbol{\beta}$ *G) + Var(E) + 2*Cov($\boldsymbol{\beta}$ *G,E)
- Recall Var(K*X) = K^2Var(X) and Cov(K*X,Y) = K*Cov(X,Y)
- So Var(T) = $\boldsymbol{\beta}^{2*}$Var(G) + Var(E) + 2$\boldsymbol{\beta}$*Cov(G,E)

Heredity (1976), **37** (1), 41-57

THE EFFECT OF CULTURAL TRANSMISSION ON CONTINUOUS VARIATION

LINDON EAVES
Department of Genetics, University of Birmingham, Birmingham B15 2TT

Received 12.xi.75

In testing these simple genotype-environment models a few equally simple environmental models have been shown to offer less satisfactory explanations of the available data. So far, there are few more subtle treatments of environmental causation which can claim to be anything better than *ad hoc* rationalisations of particular sets of data. A general quantitative theory of environmental variation is required if environmental explanations of human variation are to compete seriously with the genotype-environmental models explored so far. Failure to provide such a quantitative theory can only weaken any claim to serious attention of a purely environmental explanation of individual differences.

Heredity (1976), **37** (1), 41-57

THE EFFECT OF CULTURAL TRANSMISSION ON CONTINUOUS VARIATION

LINDON EAVES
Department of Genetics, University of Birmingham, Birmingham B15 2TT

Received 12.xi.75

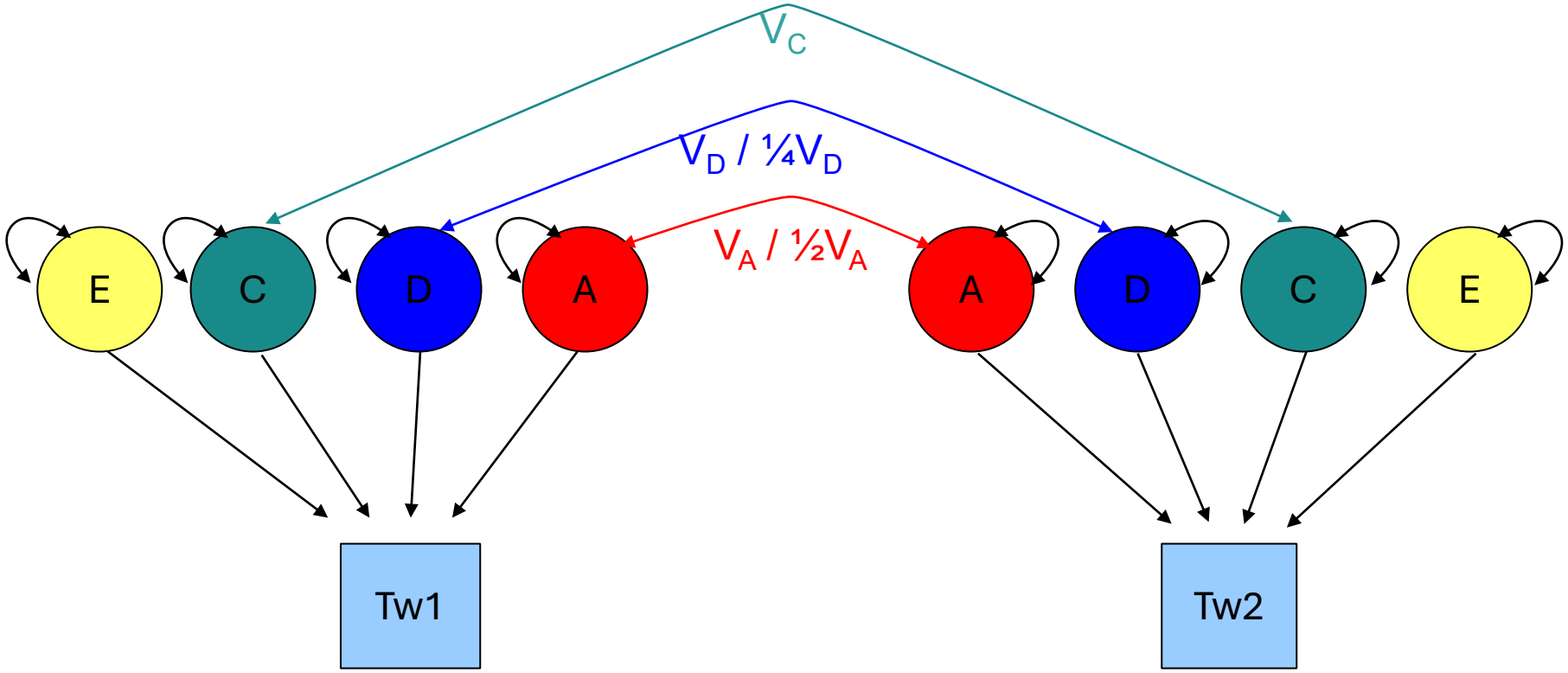
There are, however, several kinds of environmental variation which can be expressed in the form of a quantitative model. The formulation of such models is instructive because it is necessary to be quite precise about the nature of assumptions which can easily be glossed over in a merely verbal discussion of the problem. Furthermore, a precisely formulated model provides a sound basis for deciding what data need be collected in order to test the assumptions it implies and to estimate the relevant parameters of an adequate model.

Models in science

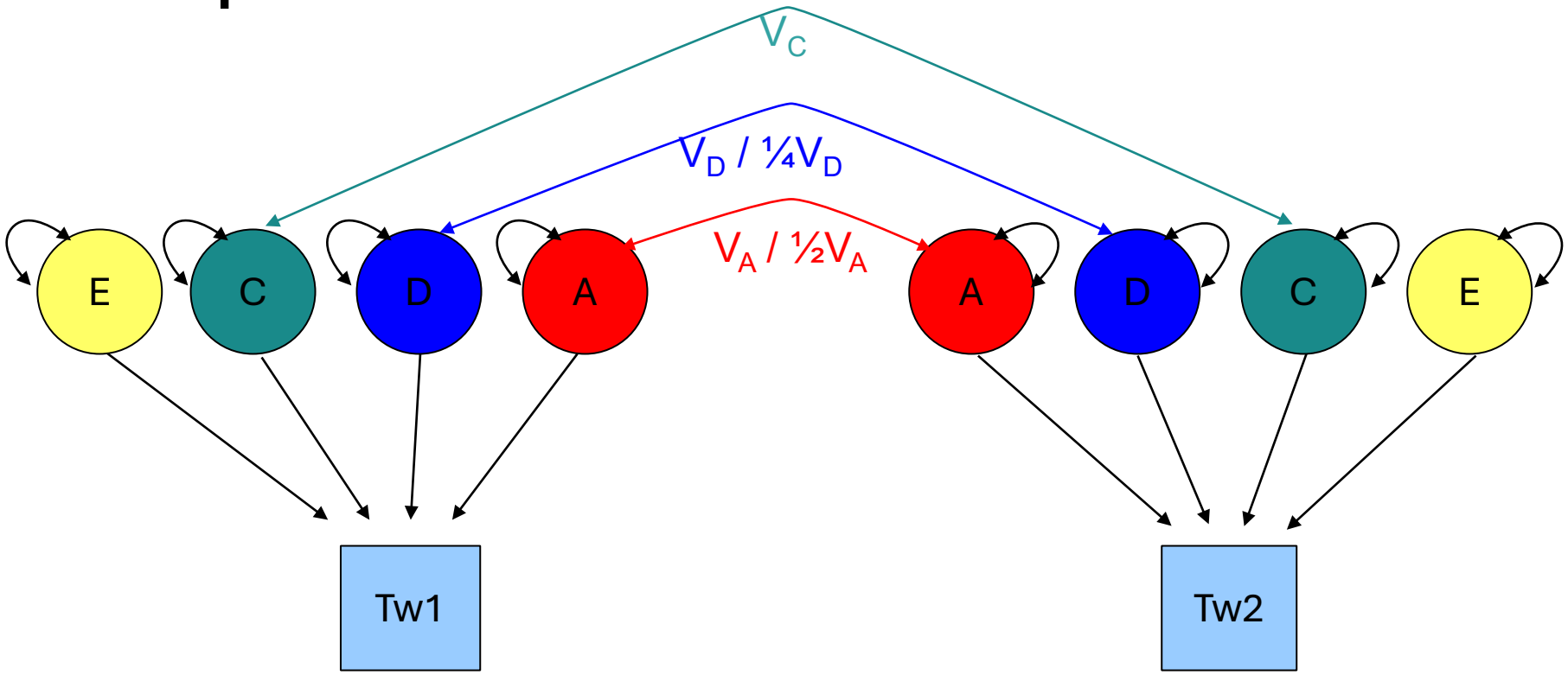
- Scientific models represent natural phenomena in a logical and simplified way, allowing for better understanding and/or prediction of the phenomena
- Models must make multiple simplifying assumptions.
- To the degree that these assumptions are unmet (do not reflect the true complexity in the real world), biases result

”All models are wrong, some are useful.”
- George Box

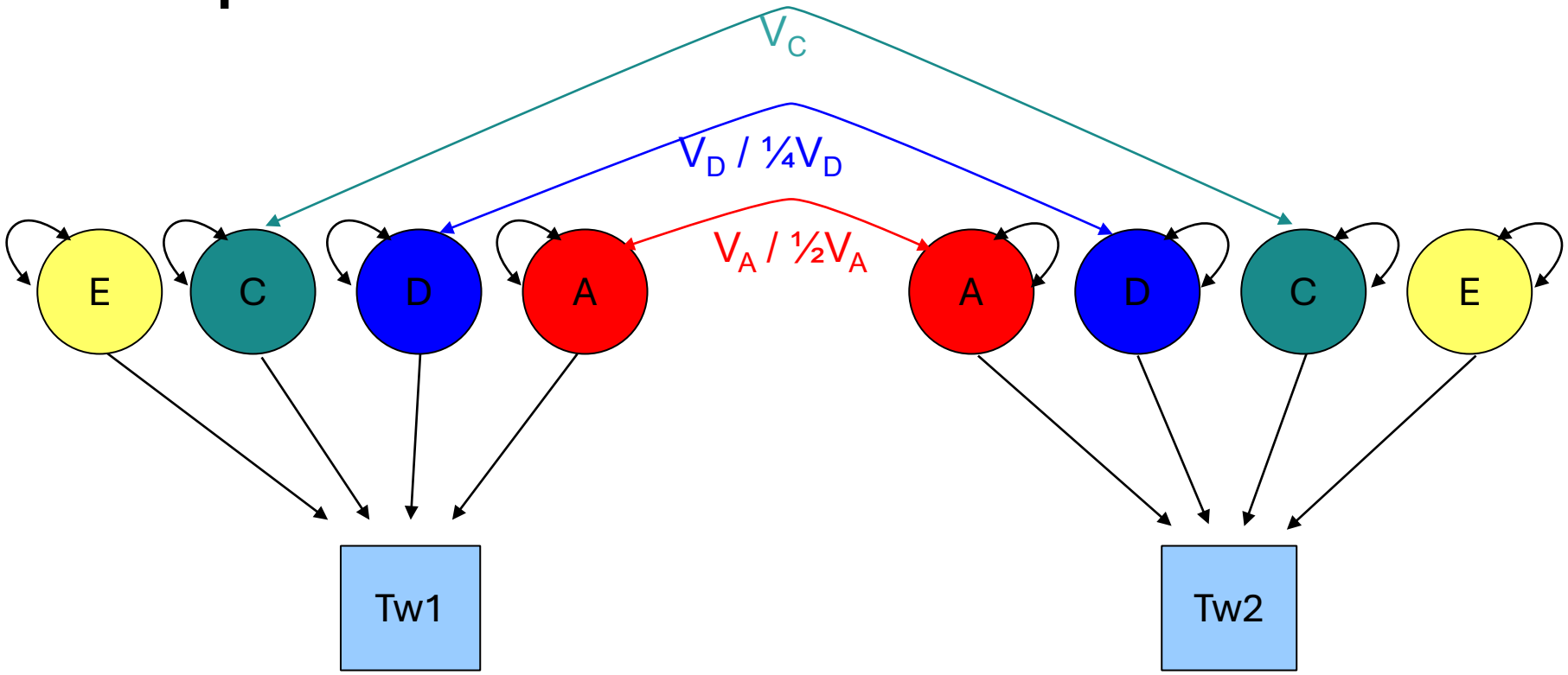
Classical twin design model



For a univariate twin model – can we estimate all these parameters?



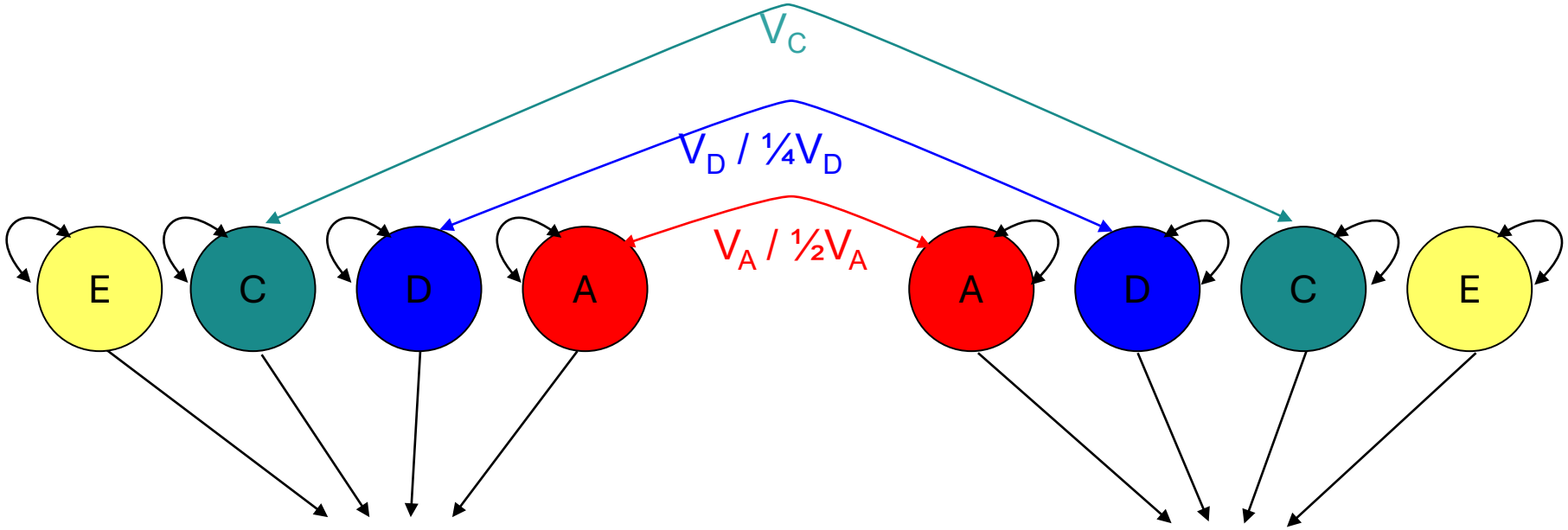
For a univariate twin model – can we estimate all these parameters? NO!



Model identification

- The number of statistics ≥ the number of parameters
  - Necessary but not sufficient
- Each parameters can be expressed as a distinct combination of statistics
  - Alternately – that the parameters are not collinear
- Algebraic identification can be performed although is considered tedious and error-prone for complex models
- Empirical detection of non-identification can be shown when different sets of parameter estimates yield identical likelihoods and those likelihoods are the maximum likelihood
- Alternatively, in OpenMx, use mxCheckIdentification(model)

So what do we do instead?

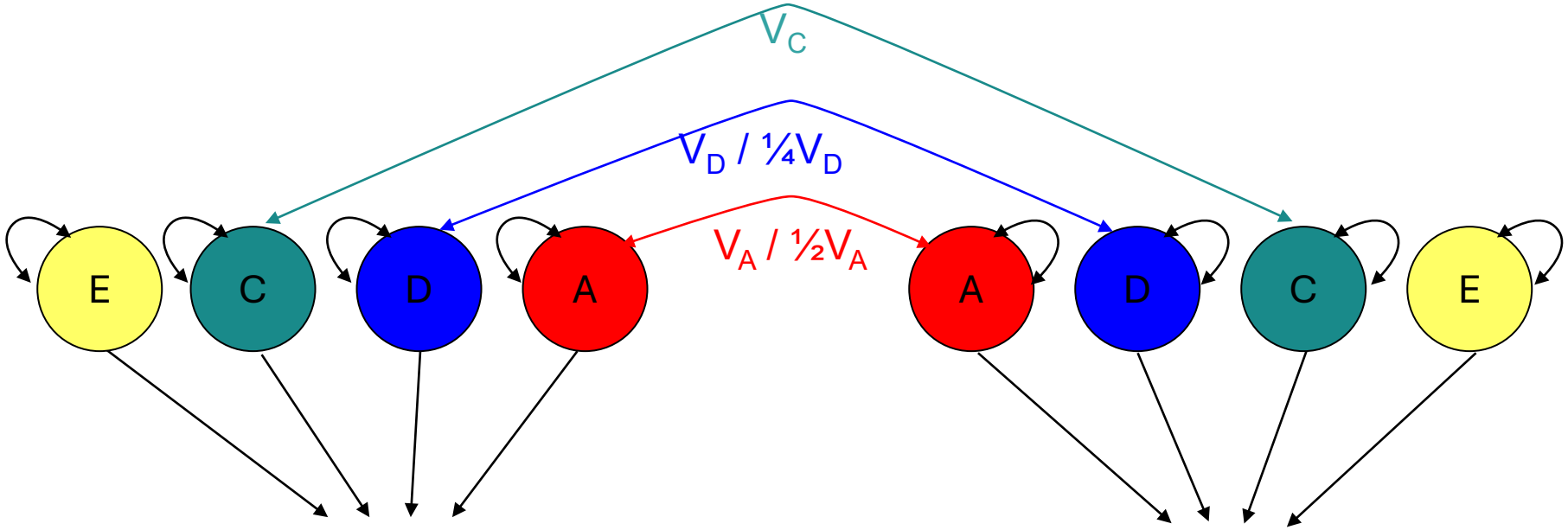


By convention, fit ACE when $\widehat{CV}_{MZ} < 2\widehat{CV}_{DZ}$

$\hat{V}_P$	
$\widehat{CV}_{MZ}$	$\hat{V}_P$

$\hat{V}_P$	
$\widehat{CV}_{DZ}$	$\hat{V}_P$

So what do we do instead?



By convention, fit ADE when $\widehat{CV}_{MZ} > 2\widehat{CV}_{DZ}$

$\hat{V}_P$	
$\widehat{CV}_{MZ}$	$\hat{V}_P$

$\hat{V}_P$	
$\widehat{CV}_{DZ}$	$\hat{V}_P$

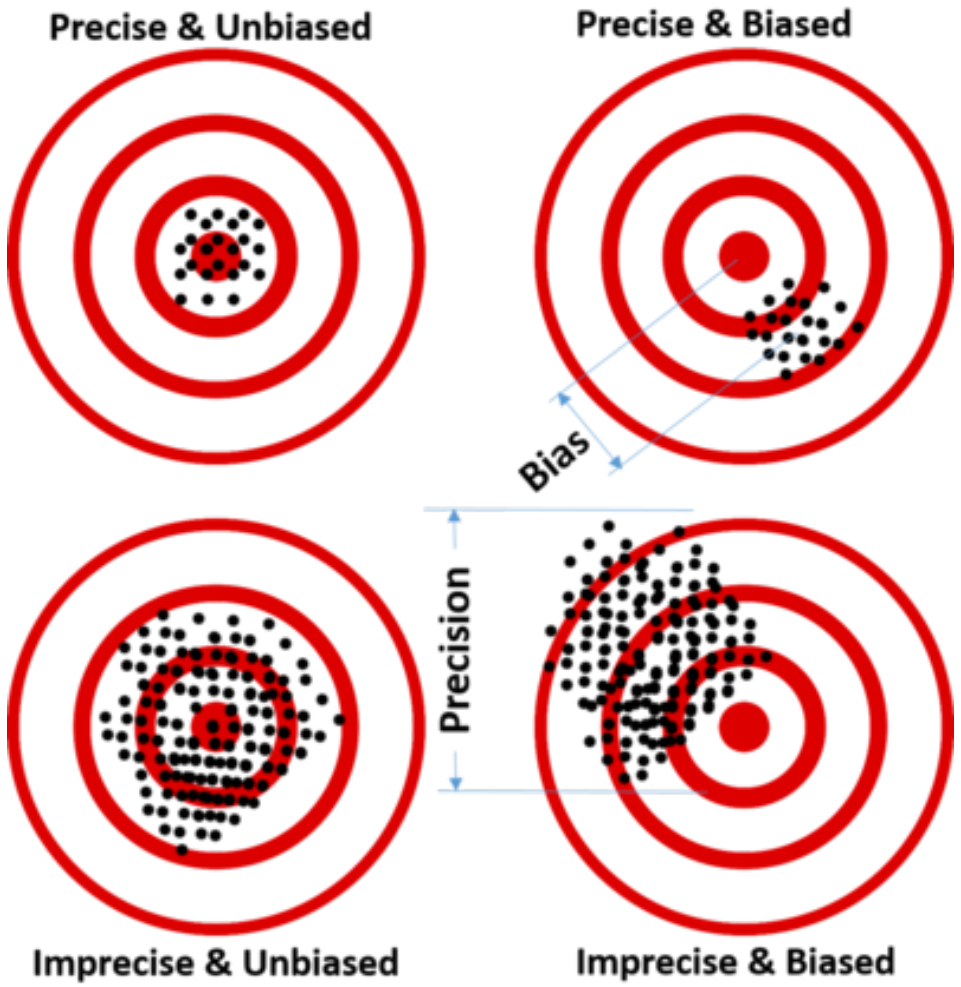
True vs. Estimated parameters

- V_A, V_C, V_D, V_E: population parameters. The **true values** (typically unknowable) in the population
- $\hat{V}_A$, $\hat{V}_C$, $\hat{V}_D$, $\hat{V}_E$: **estimated values** of V_A, V_C, V_D, and V_E
- $\hat{\theta}$ differs from θ due to:

  1) sampling variability

  2) bias (= $E[\hat{\theta}] - \theta$)

Precision and Bias - visually



Do methods have assumptions?

Do methods have assumptions? - YES

Assumptions of linear regression

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

Deriving algebraic expectations of variance component estimates

1) In an ACE model, we assume V_D=0. To get algebraic expectations of $\hat{V}_A$ and $\hat{V}_C$ in an ACE model, write down what CV_{MZ} and CV_{DZ} are assumed to be composed of:

CV_{MZ} = $V_A + V_C$

CV_{DZ} = $\frac{1}{2}V_A + V_C$

2) To get an estimate of one term (e.g., V_A), find a contrast of linear transformations of these two equations that cancel out one parameter (e.g., V_C) and isolate the other (e.g., V_A). E.g.:

$CV_{MZ} - CV_{DZ} = \frac{1}{2}V_A$. Thus $2(CV_{MZ} - CV_{DZ}) = V_A$.

Thus, an estimator of V_A:

$\hat{V}_A = 2(\widehat{CV}_{MZ} - \widehat{CV}_{DZ})$

3) Similarly, to cancel out V_A and isolate V_C:

$\hat{V}_C = 2\widehat{CV}_{DZ} - \widehat{CV}_{MZ}$

Pen & Paper Practice 1: Algebraic expectations of ADE model

Use what we just learned to derive algebraic expectations of the estimates of V_A and V_D in an ADE model (where we assume V_C=0). As a hint, in this situation, we assume V_C=0 and therefore:

$$CV_{MZ} = V_A + V_D$$
$$CV_{DZ} = \frac{1}{2}V_A + \frac{1}{4}V_D$$

To get $\hat{V}_A$, think of possible contrasts of linear transformations of these equations that cancel out V_D and isolate V_A.(and vice-versa for $\hat{V}_D$)

QUESTION1.1: What is your estimator of V_A ($\hat{V}_A$) in an ADE model?

QUESTION1.2: What is your estimator of V_D ($\hat{V}_D$) in an ADE model?

Pen & Paper Practice 1: Algebraic expectations of ADE model

QUESTION1.1: What is your estimator of V_A ($\hat{V}_A$) in an ADE model?

QUESTION1.2: What is your estimator of V_D ($\hat{V}_D$) in an ADE model?

Pen & Paper Practice 1: Algebraic expectations of ADE model

QUESTION1.1: What is your estimator of V_A ($\hat{V}_A$) in an ADE model?

CV_{MZ} = $V_A + V_D$

CV_{DZ} = $\frac{1}{2}V_A + \frac{1}{4}V_D$

$4*CV_{DZ} - CV_{MZ}$ = $2VA + VD - VA - VD = \mathbf{V_A}$

QUESTION1.2: What is your estimator of V_D ($\hat{V}_D$) in an ADE model?

CV_{MZ} = $V_A + V_D$

CV_{DZ} = $\frac{1}{2}V_A + \frac{1}{4}V_D$

$CV_{MZ} - 2* CV_{DZ} = (V_A + V_D) - 2*(\frac{1}{2}V_A + \frac{1}{4}V_D) = \frac{1}{2}V_D$

$2 *(CV_{MZ} - 2* CV_{DZ}) = \mathbf{V_D}$

Can we derive algebraic expectations of bias in estimates due to misspecification

1) We want to know what happens when we "misspecify" the model (here, when a parameter assumed to be 0 in the model is not 0). To do this, first write out one of your estimators. E.g., in an ACE model:

$$\hat{V}_A = 2(\widehat{CV}_{MZ} - \widehat{CV}_{DZ})$$

2) Consider the true compositions of parameters used in the estimators (i.e., if you got an assumption wrong). If V_D is actually non-zero, then:

$$CV_{MZ} = V_A + V_C + V_D$$
$$CV_{DZ} = \tfrac{1}{2}V_A + V_C + \tfrac{1}{4}V_D$$

3) Finally, just substitute the true compositions of CV_{MZ} and CV_{DZ} into $\widehat{CV}_{MZ}$ and $\widehat{CV}_{DZ}$ used in the estimator. Thus, for $\hat{V}_A$ in an ACE:

$$\hat{V}_A = 2*(V_A + V_D + V_C - \tfrac{1}{2}V_A - \tfrac{1}{4}V_D - V_C) = V_A + {}^{3}/_{2}V_D$$

In word: when $V_D \neq 0$ but one fits an ACE model, **$\hat{V}_A$ is biased upwards by 1.5 of whatever V_D truly is.**

4) Similarly, $\hat{V}_C = V_C - \tfrac{1}{2}V_D$: **$\hat{V}_C$ is biased down by ½ of what V_D is.**

Pen & Paper Practice 2: Deriving biases of ADE

1) Use what we just learned to derive the bias in $\hat{V}_A$ and $\hat{V}_D$ in an ADE model (where we assume V_C=0). Recall:

$$\hat{V}_A = \quad 4\widehat{CV}_{DZ} - \widehat{CV}_{MZ}$$

$$\hat{V}_D = \quad 2\widehat{CV}_{MZ} - 4\,\widehat{CV}_{DZ}$$

$$CV_{MZ} = \; V_A + \; V_D + V_C$$

$$CV_{DZ} = \; ½V_A + ¼V_D + V_C$$

2) Now just substitute the true compositions of CV_{MZ} and CV_{DZ} into $\widehat{CV}_{MZ}$ and $\widehat{CV}_{DZ}$ used in the estimator to see how our estimates are biased.

QUESTION2.1: How is $\hat{V}_A$ is biased in an ADE model when V_C (contrary to our assumption) is actually non-zero?

QUESTION2.2: How is $\hat{V}_D$ biased in an ADE model when V_C (contrary to our assumption) is actually non-zero?

Pen & Paper Practice: Deriving biases of ADE

QUESTION2.1: How is $\hat{V}_A$ biased in an ADE model when $V_C \neq 0$?

QUESTION2.2: How is $\hat{V}_D$ biased in an ADE model when $V_C \neq 0$?

Pen & Paper Practice: Deriving biases of ADE

QUESTION2.1: How is $\hat{V}_A$ biased in an ADE model when $V_C \neq 0$?

$$\hat{V}_A = 4\widehat{CV}_{DZ} - \widehat{CV}_{MZ}$$

$$\hat{V}_A = 4(\tfrac{1}{2}V_A + \tfrac{1}{4}V_D + V_C) - (V_A + V_D + V_C)$$

$$\hat{V}_A = V_A + 3V_C$$

QUESTION2.2: How is $\hat{V}_D$ biased in an ADE model when $V_C \neq 0$?

$$\hat{V}_D = 2\widehat{CV}_{MZ} - 4\,\widehat{CV}_{DZ}$$

$$\hat{V}_D = 2(V_A + V_D + V_C) - 4(\tfrac{1}{2}V_A + \tfrac{1}{4}V_D + V_C)$$

$$\hat{V}_D = V_D - 2V_C$$

Question for consideration

If the assumptions of the CTD model that either V_C or V_D is zero is violated (i.e., A, C, and D simultaneously influence phenotypic variation)… [choose all that apply]

a) the interpretation of the estimated parameters should be altered; e.g., $\hat{V}_A$ should be considered an amalgam of V_A and V_D (in ACE model) or of V_A and V_C (in ADE model)

b) there is no point in doing the analysis

c) the estimated parameter values will be biased

Question for consideration

An ADE model finds that $\hat{V}_A$ = .30 and $\hat{V}_D$ = .10. This implies that shared environmental factors do not influence the trait in question.

a) TRUE

b) FALSE

Question for consideration

We run an ADE model and find that $\hat{V}_A$ = .69 and that $\hat{V}_D$ = .05. If in truth, V_C = .10, what will the effect on the estimated parameters be? [choose all that apply]

a) $\hat{V}_A$ will be biased (too low)

b) $\hat{V}_A$ will be biased (too high)

c) $\hat{V}_D$ will be biased (too low)

d) $\hat{V}_D$ will be biased (too high)

e) there is no effect on the estimated parameters; however, by not estimating V_C (aka, fixing it to zero), we underestimated V_C

Biases in parameter estimates when V_D or V_C is 0

- In ACE Models (bias induced in setting $\hat{V}_D = 0$):

$\hat{V}_A = V_A + {}^3/_2\, V_D$

- $\hat{V}_C = V_C - {}^1/_2\, V_D$

- In ADE Models (bias induced in setting $\hat{V}_C = 0$):

$\hat{V}_A = V_A + 3V_C$

- $\hat{V}_D = V_D - 2V_C$

- Thus, V_A is typically over-estimated and V_C and V_D under-estimated.
- However, things are more complicated when one considers the possibility of epistasis, assortative mating, etc.

Question for consideration

What are the *typical* assumptions of a classical twin model? [choose all that apply]

a) only genetic factors cause MZ twins to be more correlated than DZ twins

b) either V_D or V_C is zero

c) no epistasis

d) no assortative mating

e) no gene-environment interactions or correlations

What are the typical effects of violations of assumptions in the CTD?

a) Only genetic factors cause MZ twins to be more correlated than DZ twins:
$\hat{V}_A$ & $\hat{V}_D$ overestimated and $\hat{V}_C$ underestimated

b) Either V_D or V_C is zero:
$\hat{V}_A$ overestimated and $\hat{V}_D$ & $\hat{V}_C$ underestimated

c) No epistasis:
$\hat{V}_D$ or $\hat{V}_A$ overestimated and $\hat{V}_C$ underestimated

d) No assortative mating:
$\hat{V}_A$ and/or $\hat{V}_D$ underestimated and $\hat{V}_C$ overestimated

e) No gene-environment interactions or correlations:
AxC: $\hat{V}_A$ overestimated
AxE: $\hat{V}_E$ overestimated
passive Cov(A,C): $\hat{V}_C$ overestimate

Conclusions

- All models require assumptions. Generally, the more these assumptions are violated, the more estimates are biased
- Understanding biases allows you to understand how to interpret estimates with the proper nuance
- In all models, including the CTD, be cautious of reifying parameter estimates!
    - $\hat{V}_A$ is amalgam of mostly V_A but also V_D & V_C.
    - $\hat{V}_C$ & $\hat{V}_D$ may often be underestimates
    - Interpret $\hat{V}_D$ as a (potentially downwardly biased) estimate of V_{NA}
    - $\hat{V}_A/\hat{V}_P$ (in ACE) or $(\hat{V}_A+\hat{V}_D)/\hat{V}_P$ (in ADE) are decent estimates of broad sense h^2.