Class Introduction and Overview; Review of Matrix Algebra and the Multivariate Normal Distribution Introduction to Mplus

Latent Trait Measurement and Structural Equation Models
Lecture #1 – January 9, 2013

Today's Class

- Introduction and overview of the course
  - Syllabus information
- Review of prerequisites
  - Matrix algebra
  - Multivariate normal distribution
- Introduction to Mplus

Today's Data Set

- To introduce and motivate SEM, and to review some prerequisites, we will make use of an example data set

- Data come from a sample of 200 (79 men and 121 women) participants in a study of the role of self esteem and motivation on achievement test scores

- Participants responded to three tests/surveys:
  - 20-item achievement test (each item was scored right/wrong)
  - 5-item motivation survey (each item used a 9-point Likert scale for responses; 1-9 in integers)
  - 5-item self esteem survey (each item used a 9-point Likert scale for responses; 1-9 in integers)

- The researchers were interested in the effects of motivation and self esteem on achievement

Data File Setup

- The data file (exampledata.xls) has the following variables:
  - ID – identification number for each respondent
  - AchievementScore – Total score for achievement items
  - SelfEsteemScore – Total score for self esteem items
  - MotivationScore – Total score for motivation items
  - SelfEsteem1-SelfEsteem5 – Self esteem item responses
  - Motivation1-Motivation5 – Motivation item responses
  - Achievement1-Achievement20 – Achievement item responses

- Note: In order to use this file with Mplus, you must:
  - Save it as a comma-delimited file (.csv)
  - Remove the first row containing the variable names

MOTIVATION FOR LEARNING SEM

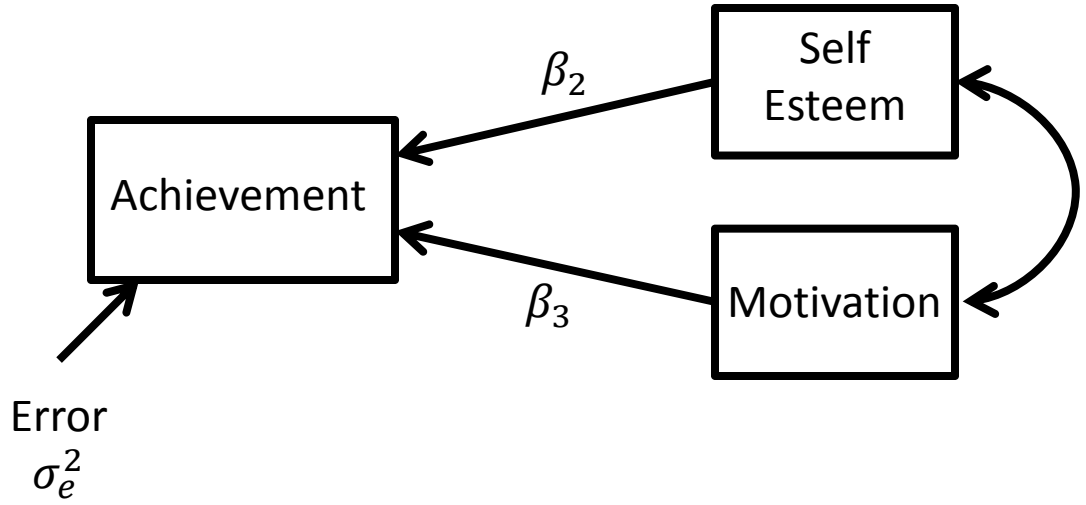
The Answer…Don't Use Aggregates – USE SEM

- Structural Equation Modeling seeks to determine the relationship between:
  - Latent constructs only
    - Latent our example: Achievement, Motivation, and Self Esteem
  - Latent and observed constructs
    - From our example: how does gender factor into the model?
  - Complex relationships between latent constructs and observed variables
    - Does motivation mediate the relationship between self esteem and achievement?
    - Do either/both mediate the relationship between gender and achievement?

- SEM is a generalization of linear modeling using observed and latent (sometimes called random) variables
  - I tend to think of SEM as a part of a bigger picture…you will see that SEM people think everything is part of SEM

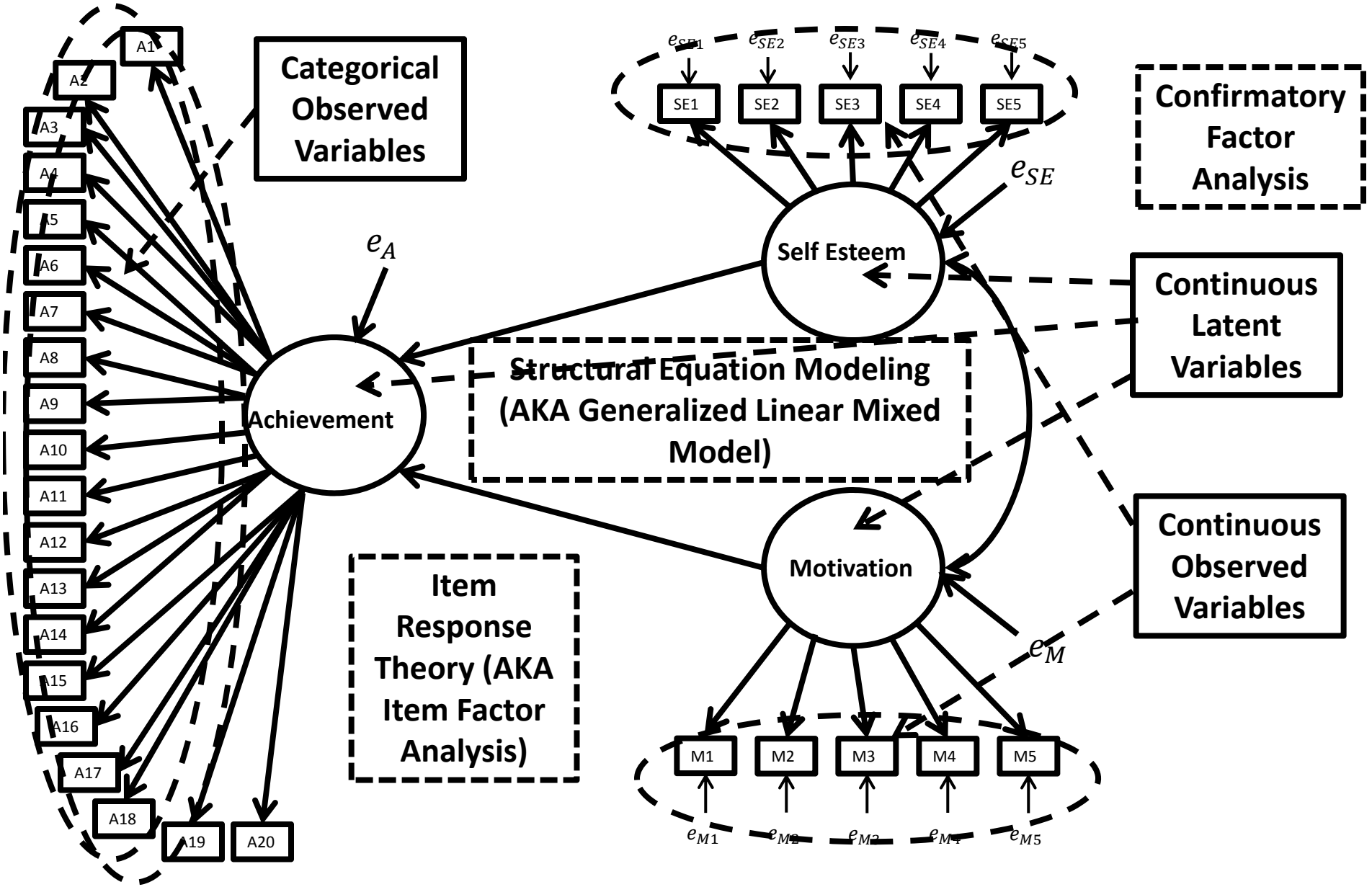
Path Diagram of Our Regression Example

- A common way of depicting models in SEM is with a path diagram :: a pictorial representation of the statistical model
  - Observed variables: Squares
  - Latent variables: Circles
  - Direct effects: Arrows with one head
  - Indirect effects: Arrows with two heads



Path Diagram of Regression Result

A More Accurate Path Diagram



The (Really) Big Picture

- Statistical distributions are what drive the process
  - Each distribution is described by a set of parameters
  - Think of the normal distribution (mean and variance)

- Each of the lines represents model parameters
  - The statistical distribution of the boxes and circles are described by the model parameters

- Model parameters provide constraints to the statistical distribution parameters
  - Reduce complexity of model
  - Provide for meaningful inference

- A model is bound by distributions assumed and, hence, the number of possible parameters
  - We will learn statistics and path models
    - Both are needed to be good at SEM

MATRIX ALGEBRA

Introduction and Motivation for Matrix Algebra

- Structural equation modeling and nearly all other multivariate statistical techniques are described with matrix algebra
  - It is the language of modern statistics

- When new methods are developed, the first published work typically involves matrices
  - It makes technical writing more concise – formulae are smaller

- Have you seen:
  - $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ (from the general linear model)
  - $\boldsymbol{\Lambda\Phi\Lambda}^T + \boldsymbol{\Psi}$ (from confirmatory factor analysis)

- **Useful tip:** matrix algebra is a great way to get out of conversations and other awkward moments

Why Learn Matrix Algebra

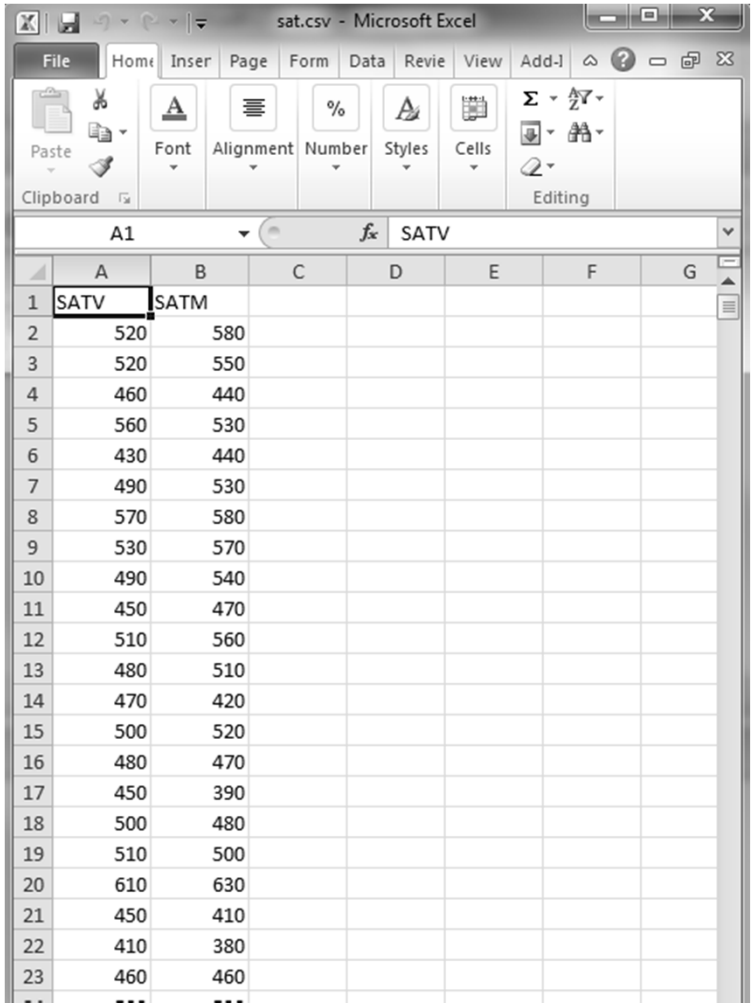
- Matrix algebra can seem very abstract from the purposes of this class (and statistics in general)

- Learning matrix algebra is important for:
  - Understanding how statistical methods work
    - And when to use them (or not use them)
  - Understanding what statistical methods mean
  - Reading and writing results from new statistical methods

- Today's class is the first lecture of learning the language of structural equation modeling

The Data…

In Excel:



	A	B	C	D	E	F	G
1	SATV	SATM					
2	520	580					
3	520	550					
4	460	440					
5	560	530					
6	430	440					
7	490	530					
8	570	580					
9	530	570					
10	490	540					
11	450	470					
12	510	560					
13	480	510					
14	470	420					
15	500	520					
16	480	470					
17	450	390					
18	500	480					
19	510	500					
20	610	630					
21	450	410					
22	410	380					
23	460	460					

THE BASICS: DEFINITIONS OF MATRICES, VECTORS, AND SCALARS

Matrices

- A matrix is a rectangular array of data
  - Used for storing numbers

- Matrices can have unlimited dimensions
  - For our purposes all matrices will have two dimensions:
    - Row
    - Columns

- Matrices are symbolized by **boldface** font in text, typically with capital letters
  - Size (r rows x c columns)

$$\mathbf{X} = \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix}_{(1000\ x\ 2)}$$

Vectors

- A vector is a matrix where one dimension is equal to size 1
  - Column vector: a matrix of size $r \; x \; 1$

$$\boldsymbol{x}_{\cdot 1} = \begin{bmatrix} 520 \\ 520 \\ \vdots \\ 540 \end{bmatrix}_{1000 \; x \; 1}$$

  - Row vector: a matrix of size $1 \; x \; c$

$$\boldsymbol{x}_{1\cdot} = [520 \quad 580]_{1 \; x \; 2}$$

- Vectors are typically written in **boldface** font text, usually with lowercase letters

Scalars

- A scalar is just a single number (as we have known before)

- The name scalar is important: the number “scales” a vector – it can make a vector “longer” or “shorter”

Matrix Elements

- A matrix (or vector) is composed of a set of elements
  - Each element is denoted by its position in the matrix (row and column)

- For our matrix of data $\mathbf{X}$ (size 1000 rows and 2 columns), each element is denoted by:

$$x_{ij}$$

  - The first subscript is the index for the rows: i = 1,…,r (= 1000)
  - The second subscript is the index for the columns: j = 1,…,c (= 2)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{1000,1} & x_{1000,2} \end{bmatrix}_{(1000\ x\ 2)}$$

Matrix Transpose

- The transpose of a matrix is a reorganization of the matrix by switching the indices for the rows and columns

$$\mathbf{X} = \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix}_{(1000\ x\ 2)}$$

$$\mathbf{X}^T = \begin{bmatrix} 520 & 520 & \cdots & 540 \\ 580 & 550 & \cdots & 660 \end{bmatrix}_{(2\ x\ 1000)}$$

- An element x_{ij} in the original matrix $\mathbf{X}$ is now x_{ji} in the transposed matrix $\mathbf{X}^T$

Types of Matrices

- **Square Matrix:** A matrix that has the same number of rows and columns
  - Correlation/covariance matrices are square matrices

- **Diagonal Matrix:** A diagonal matrix is a square matrix with non-zero diagonal elements ($x_{ij} \neq 0 \text{ for } i = j$) and zeros on the off-diagonal elements ($x_{ij} = 0 \text{ for } i \neq j$):

$$\mathbf{A} = \begin{bmatrix} 2.759 & 0 & 0 \\ 0 & 1.643 & 0 \\ 0 & 0 & 0.879 \end{bmatrix}$$

  - We will use diagonal matrices to form correlation matrices

- **Symmetric Matrix:** A symmetric matrix is a square matrix where all elements are reflected across the diagonal ($a_{ij} = a_{ji}$)
  - Correlation and covariance matrices are symmetric matrices

MATRIX ALGEBRA

Moving from Vectors to Matrices

- A matrix can be thought of as a collection of vectors
  - Matrix operations are vector operations on steroids
- Matrix algebra defines a set of operations and entities on matrices
  - I will present a version meant to mirror your previous algebra experiences
- Definitions:
  - Identity matrix
  - Zero vector
  - Ones vector
- Basic Operations:
  - Addition
  - Subtraction
  - Multiplication
  - “Division”

Matrix Addition and Subtraction

- Matrix addition and subtraction are much like vector addition/subtraction

- Rules:
  - Matrices must be the same size (rows and columns)

- Method:
  - The new matrix is constructed of element-by-element addition/subtraction of the previous matrices

- Order:
  - The order of the matrices (pre- and post-) does not matter

Matrix Addition/Subtraction

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \\ a_{31} + b_{31} & a_{32} + b_{32} \\ a_{41} + b_{41} & a_{42} + b_{42} \end{bmatrix} \qquad \mathbf{A} - \mathbf{B} = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} \\ a_{21} - b_{21} & a_{22} - b_{22} \\ a_{31} - b_{31} & a_{32} - b_{32} \\ a_{41} - b_{41} & a_{42} - b_{42} \end{bmatrix}$$

More Matrix Addition

- Matrix addition…using our data matrix:

$$\mathbf{X} = \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix}_{(1000\ x\ 2)}$$

$$\mathbf{X} + \mathbf{X} = \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix} + \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix} = \begin{bmatrix} 520+520 & 580+580 \\ 520+520 & 550+550 \\ \vdots & \vdots \\ 540+540 & 660+660 \end{bmatrix}$$

$$= \begin{bmatrix} 1040 & 1160 \\ 1040 & 1100 \\ \vdots & \vdots \\ 1080 & 1320 \end{bmatrix}$$

Matrix Multiplication

- Matrix multiplication is a bit more complicated
  - The new matrix may be a different size from either of the two multiplying matrices

$$\mathbf{A}_{(r\ x\ c)}\mathbf{B}_{(c\ x\ k)} = \mathbf{C}_{(r\ x\ k)}$$

- Rules:
  - Pre-multiplying matrix must have number of columns equal to the number of rows of the post-multiplying matrix

- Method:
  - The elements of the new matrix consist of the inner (dot) product of the row vectors of the pre-multiplying matrix and the column vectors of the post-multiplying matrix

- Order:
  - The order of the matrices (pre- and post-) matters

Matrix Multiplication

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} & a_{31}b_{13} + a_{32}b_{23} \\ a_{41}b_{11} + a_{42}b_{21} & a_{41}b_{12} + a_{42}b_{22} & a_{41}b_{13} + a_{42}b_{23} \end{bmatrix}$$

Multiplication in Statistics

- Many statistical formulae with summation can be re-expressed with matrices

- A common matrix multiplication form is: $\mathbf{X}^T\mathbf{X}$
  - Diagonal elements: $\sum_{i=1}^{N} X_i^2$
  - Off-diagonal elements: $\sum_{i=1}^{N} X_{ia}X_{ib}$

- For our SAT example:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{N} SATV_i^2 & \sum_{i=1}^{N} SATV_i SATM_i \\ \sum_{i=1}^{N} SATV_i SATM_i & \sum_{i=1}^{N} SATM_i^2 \end{bmatrix} = \begin{bmatrix} 251{,}797{,}800 & 251{,}928{,}400 \\ 251{,}928{,}400 & 254{,}862{,}700 \end{bmatrix}$$

Matrix Multiplication…with Numbers

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 520 & 520 & \cdots & 540 \\ 580 & 550 & \cdots & 660 \end{bmatrix} \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix}$$

$$= \begin{bmatrix} 520 * 520 + 520 * 520 + \cdots + 540 * 540 & 520 * 580 + 520 * 550 + 540 * 660 \\ 580 * 520 + 550 * 520 + \cdots + 660 * 540 & 580 * 580 + 550 * 550 + 660 * 660 \end{bmatrix}$$

$$= \begin{bmatrix} 251{,}797{,}800 & 251{,}928{,}400 \\ 251{,}928{,}400 & 254{,}862{,}700 \end{bmatrix}$$

Identity Matrix

- The identity matrix is a matrix that, when pre- or post- multiplied by another matrix results in the original matrix:

$$\mathbf{AI} = \mathbf{A}$$
$$\mathbf{IA} = \mathbf{A}$$

- The identity matrix is a square matrix that has:
  - Diagonal elements = 1
  - Off-diagonal elements = 0

$$I_{(3\ x\ 3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Zero Vector

- The zero vector is a column vector of zeros

$$\mathbf{0}_{(3\ x\ 1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

- When pre- or post- multiplied the result is the zero vector:

$$\mathbf{A0} = \mathbf{0}$$
$$\mathbf{0A} = \mathbf{0}$$

Ones Vector

- A ones vector is a column vector of 1s:

$$\mathbf{1}_{(3\ x\ 1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

- The ones vector is useful for calculating statistical terms, such as the mean vector and the covariance matrix
  - Next class we will discuss what these matrices are, how we compute them, and what the mean

Matrix "Division": The Inverse Matrix

- Division from algebra:
  - First: $\frac{a}{b} = \frac{1}{b}a = b^{-1}a$
  - Second: $\frac{a}{a} = 1$

- "Division" in matrices serves a similar role
  - For square and symmetric matrices, an inverse matrix is a matrix that when pre- or post- multiplied with another matrix produces the identity matrix:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

- Calculation of the matrix inverse is complicated
  - Even computers have a tough time

- Not all matrices can be inverted
  - Non-invertable matrices are called singular matrices
    - In statistics, singular matrices are commonly caused by linear dependencies

The Inverse

- **<u>In data:</u>** the inverse shows up constantly in statistics
  - Models which assume some type of (multivariate) normality need an inverse covariance matrix

- Using our SAT example
  - Our data matrix was size (1000 x 2), which is not invertible
  - However $\mathbf{X}^T\mathbf{X}$ was size (2 x 2) – square, and symmetric

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 251{,}797{,}800 & 251{,}928{,}400 \\ 251{,}928{,}400 & 254{,}862{,}700 \end{bmatrix}$$

  - The inverse is:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 3.61E-7 & -3.57E-7 \\ -3.57E-7 & 3.56E-7 \end{bmatrix}$$

Matrix Algebra Operations

- $(\mathbf{A}+\mathbf{B})+\mathbf{C}=$
  $\mathbf{A}+(\mathbf{B}+\mathbf{C})$
- $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$
- $c(\mathbf{A}+\mathbf{B})=c\mathbf{A}+c\mathbf{B}$
- $(c+d)\mathbf{A}=c\mathbf{A}+d\mathbf{A}$
- $(\mathbf{A}+\mathbf{B})^T=\mathbf{A}^T+\mathbf{B}^T$
- $(cd)\mathbf{A}=c(d\mathbf{A})$
- $(c\mathbf{A})^T=c\mathbf{A}^T$
- $c(\mathbf{AB})=(c\mathbf{A})\mathbf{B}$
- $\mathbf{A}(\mathbf{BC})=(\mathbf{AB})\mathbf{C}$
- $\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC}$
- $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$
- For x_j such that Ax_j exists:

$$\sum_{j=1}^{N}\mathbf{A}\mathbf{x}_j=\mathbf{A}\sum_{j=1}^{N}\mathbf{x}_j$$

$$\sum_{j=1}^{N}(\mathbf{A}\mathbf{x}_j)(\mathbf{A}\mathbf{x}_j)^T=$$

$$\mathbf{A}\left(\sum_{j=1}^{N}\mathbf{x}_j\mathbf{x}_j^T\right)\mathbf{A}^T$$

UNIVARIATE STATISTICAL DISTRIBUTIONS

Univariate Normal Distribution

- For a continuous random variable x (ranging from $-\infty$ to ∞) the univariate normal distribution function is:

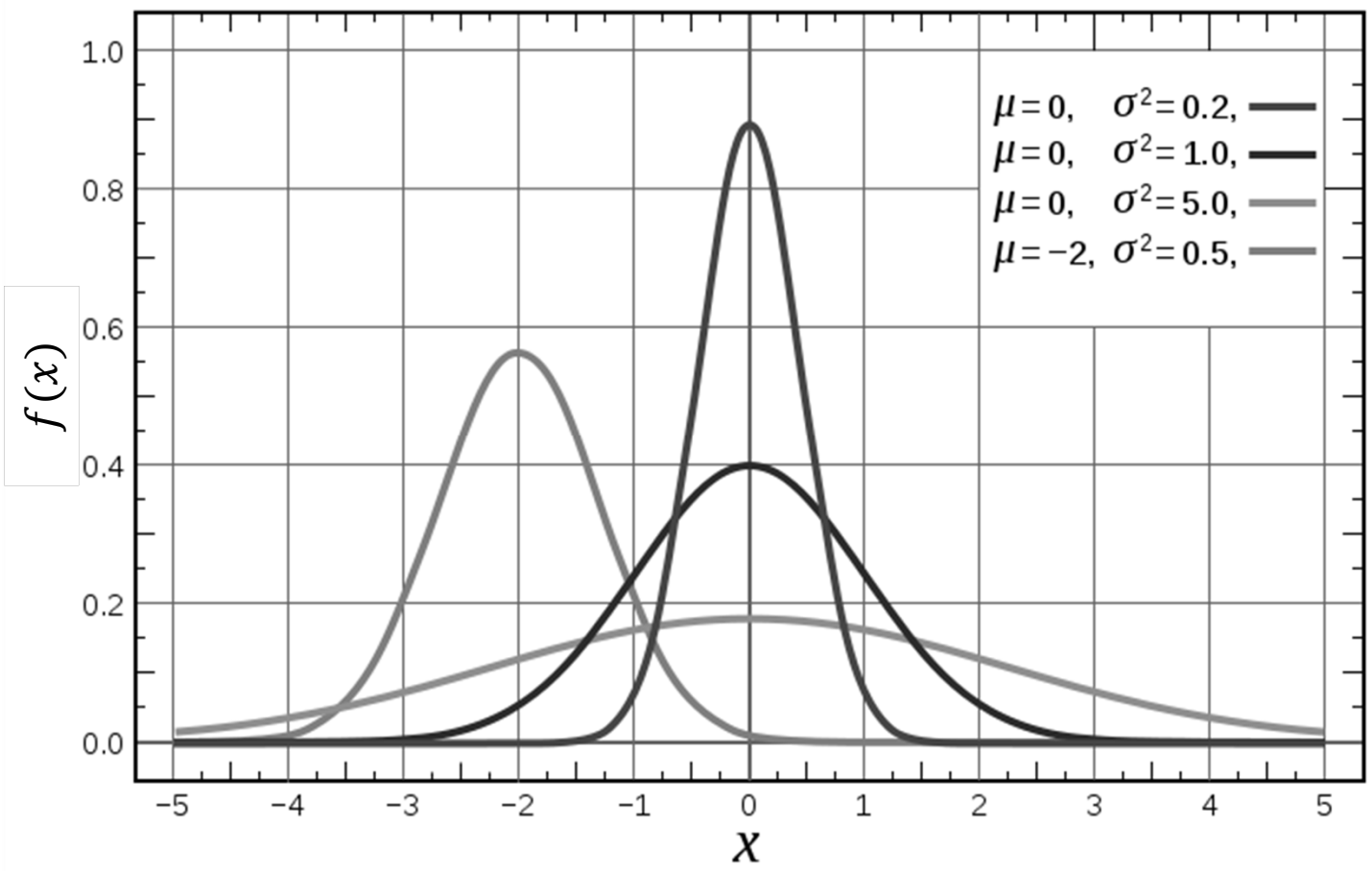
$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

- The shape of the distribution is governed by two parameter:
  - The mean μ_x
  - The variance σ_x^2

- The skewness (lean) and kurtosis (peakedness) are fixed

- Standard notation for normal distributions is $X \sim N(\mu_x, \sigma_x^2)$

Univariate Normal Distribution



For any value of x, $f(x)$ gives the height of the curve (relative frequency)

Uses of Distributions

- Statistical models make distributional assumptions on various parameters and/or parts of data

- These assumptions govern:
  - How models are estimated
  - How inferences are made
  - How missing data may be imputed

- If data do not follow an assumed distribution, inferences may be inaccurate
  - Sometimes a problem, other times not so much

- Therefore, it can be helpful to check distributional assumptions prior to (or while) running statistical analyses

BIVARIATE STATISTICS AND DISTRIBUTIONS

Bivariate Statistics

- Up to this point, we have focused on only one of our variables: height
  - Looked at its **marginal distribution** (the distribution of height independent of that of weight)
  - Could have looked at weight, marginally

- Structural equation modeling is about exploring joint distributions of multiple variables
  - How sets of variables relate to each other

- As such, we will now look at the joint distributions of two variables (x_1, x_2) or in matrix form: $\mathbf{X}$ (size N x 2)
  - Beginning with two, then moving to anything more than two

Multiple Means: The Mean Vector

- We can use a vector to describe the set of means for our data

$$\bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^T\mathbf{1} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

  - Here $\mathbf{1}$ is a N x 1 vector of 1s
  - The resulting mean vector is a *p* x 1 vector of means

- For our data:

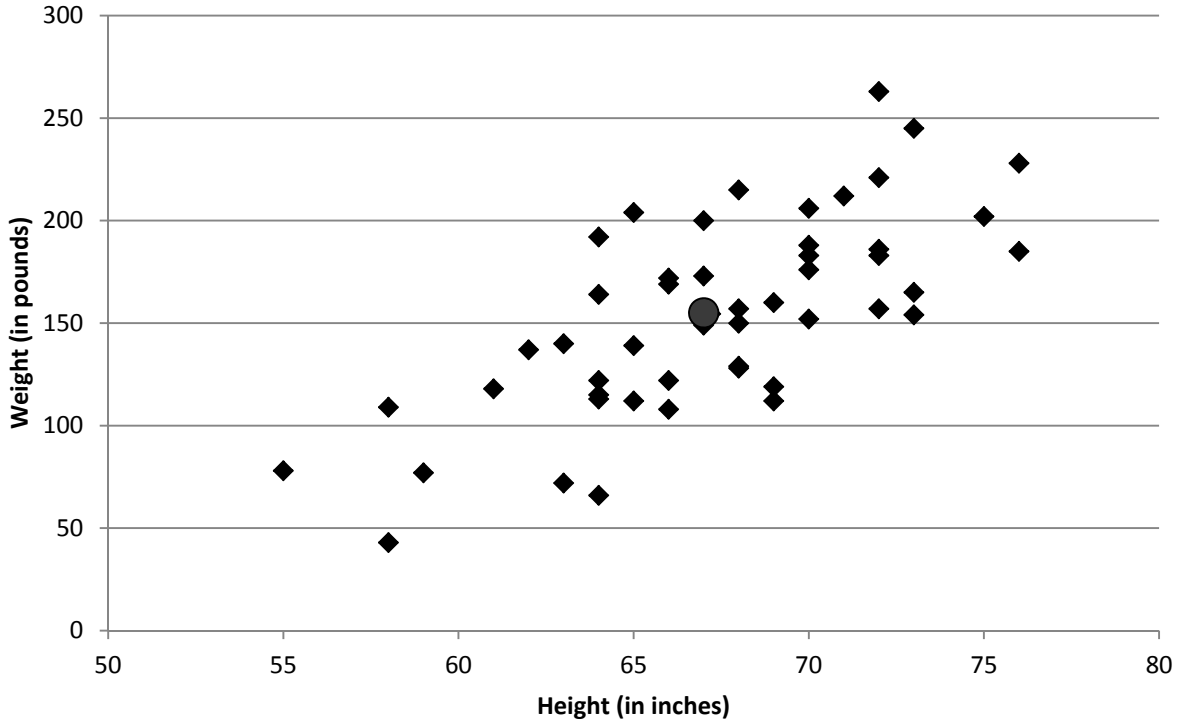
$$\bar{\mathbf{x}} = \begin{bmatrix} 67.2 \\ 154.5 \end{bmatrix} = \begin{bmatrix} \bar{x}_{height} \\ \bar{x}_{weight} \end{bmatrix}$$

- From Mplus:

```
Means
    HEIGHT            67.280      0.653    103.029      0.000
    WEIGHT           154.500      6.656     23.213      0.000
```

Mean Vector: Graphically

- The mean vector is the center of the distribution of both variables



Covariance of a Pair of Variables

- The covariance is a measure of the relatedness
  - Expressed in the product of the units of the two

$$s_{x_1 x_2} = \frac{1}{N}\sum_{i=1}^{N}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

  - The covariance between height and weight was 155.4 (in inch-pounds)
  - The denominator N is the ML version – unbiased is N-1

- Because the units of the covariance are difficult to understand, we more commonly describe association (correlation) between two variables with correlation
  - Covariance divided by the product of each variable's standard deviation

Covariance in Mplus

- Because Mplus defaults to the multivariate normal distribution, covariance is the type of parameter natively estimated by the program
  - Additionally, the MODEL RESULTS are maximum likelihood estimates (MLEs, see next week's lecture)
    - This means the denominator is N, not N-1

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HEIGHT WITH				
WEIGHT	155.400	37.783	4.113	0.000
Means				
HEIGHT	67.280	0.653	103.029	0.000
WEIGHT	154.500	6.656	23.213	0.000
Variances				
HEIGHT	21.322	4.264	5.000	0.000
WEIGHT	2215.007	443.001	5.000	0.000

Correlation of a Pair of Variables

- Correlation is covariance divided by the product of the standard deviation of each variable:

$$r_{x_1x_2} = \frac{s_{x_1x_2}}{\sqrt{s_{x_1}^2}\sqrt{s_{x_2}^2}}$$

  - The correlation between height and weight was 0.72

- Correlation is unitless – it only ranges between -1 and 1
  - If x_1 and x_2 had variances of 1, the covariance between them would be a correlation
    - Covariance of standardized variables = correlation

Covariance and Correlation in Matrices

- The covariance matrix (for any number of variables *p*) is found by:

$$\mathbf{S} = \frac{1}{N}(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T)^T(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T) = \begin{bmatrix} s_{x_1}^2 & \cdots & s_{x_1x_p} \\ \vdots & \ddots & \vdots \\ s_{x_1x_p} & \cdots & s_{x_p}^2 \end{bmatrix}$$

- If we take the SDs (the square root of the diagonal of the covariance matrix) and put them into a diagonal matrix $\mathbf{D}$, the correlation matrix is found by:

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1} = \begin{bmatrix} \frac{s_{x_1}^2}{\sqrt{s_{x_1}^2}\sqrt{s_{x_1}^2}} & \cdots & \frac{s_{x_1x_p}}{\sqrt{s_{x_1}^2}\sqrt{s_{x_p}^2}} \\ \vdots & \ddots & \vdots \\ \frac{s_{x_1x_p}}{\sqrt{s_{x_1}^2}\sqrt{s_{x_p}^2}} & \cdots & \frac{s_{x_p}^2}{\sqrt{s_{x_p}^2}\sqrt{s_{x_p}^2}} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & r_{x_1x_p} \\ \vdots & \ddots & \vdots \\ r_{x_1x_p} & \cdots & 1 \end{bmatrix}$$

Example Covariance Matrix

- For our data, the covariance matrix was:

$$\mathbf{S} = \begin{bmatrix} 21.322 & 154.5 \\ 154.5 & 2{,}215.007 \end{bmatrix}$$

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HEIGHT WITH				
WEIGHT	155.400	37.783	4.113	0.000
Means				
HEIGHT	67.280	0.653	103.029	0.000
WEIGHT	154.500	6.656	23.213	0.000
Variances				
HEIGHT	21.322	4.264	5.000	0.000
WEIGHT	2215.007	443.001	5.000	0.000

- The diagonal matrix $\mathbf{D}$ was:

$$\mathbf{D} = \begin{bmatrix} \sqrt{21.322} & 0 \\ 0 & \sqrt{2{,}215.007} \end{bmatrix}$$

- The correlation matrix $\mathbf{R}$ was:

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{21.322}} & 0 \\ 0 & \frac{1}{\sqrt{2{,}215.007}} \end{bmatrix} \begin{bmatrix} 21.322 & 154.5 \\ 154.5 & 2{,}215.007 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{21.322}} & 0 \\ 0 & \frac{1}{\sqrt{2{,}215.007}} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.715 \\ 0.715 & 1.000 \end{bmatrix}$$

Correlation and Covariance in Mplus

- Mplus does not directly estimate correlation
  - But you can obtain it in two ways with the output section:

SAMPSTAT

```
SAMPLE STATISTICS


           Means
              HEIGHT         WEIGHT
              _______        _______
    1         67.280         154.500


           Covariances
              HEIGHT         WEIGHT
              _______        _______
HEIGHT        21.322
WEIGHT       155.400       2215.010


           Correlations
              HEIGHT         WEIGHT
              _______        _______
HEIGHT         1.000
WEIGHT         0.715          1.000
```

STANDARDIZED

```
STANDARDIZED MODEL RESULTS


STDYX Standardization

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 HEIGHT   WITH
    WEIGHT             0.715      0.069     10.347      0.000

 Means
    HEIGHT            14.571      1.464      9.953      0.000
    WEIGHT             3.283      0.357      9.184      0.000

 Variances
    HEIGHT             1.000      0.000    999.000    999.000
    WEIGHT             1.000      0.000    999.000    999.000
```

Cov is $\frac{COV(X,Y)}{SD(X)SD(Y)}$

Mean is $\frac{MEAN(X)}{SD(X)}$

Generalized Variance

- The determinant of the sample covariance matrix is called the generalized variance

$$\text{Generalized Sample Variance} = |\mathbf{S}|$$

- It is a measure of spread across all variables
  - Reflecting how much overlap (covariance) in variables occurs
  - Amount of overlap reduces the generalized sample variance

- The generalized sample variance is:
  - Largest when variables are uncorrelated
  - Zero when variables form a linear dependency

- **<u>In data:</u>**
  - The generalized variance is seldom used descriptively, but shows up more frequently in maximum likelihood functions

Total Sample Variance

- The total sample variance is the sum of the variances of each variable in the sample
  - The sum of the diagonal elements of the sample covariance matrix
  - The trace of the sample covariance matrix

$$Total\ Sample\ Variance = \sum_{i=1}^{p} s_{x_i}^2 = \text{tr}\,\mathbf{S}$$

- The total sample variance does not take into consideration the covariances among the variables
  - Will not equal zero if linearly dependency exists

- **<u>In data:</u>**
  - The total sample variance is commonly used as the denominator (target) when calculating variance accounted for measures

BIVARIATE NORMAL DISTRIBUTION

Bivariate Normal Distribution

- The bivariate normal distribution is a statistical distribution for two variables
  - Both variable is normally distributed marginally (by itself)
  - Together, they form a bivariate normal distribution

- The bivariate normal density provides the relatively frequency of observing any **pair** of observations, $\mathbf{x}_i = [x_{i1} \quad x_{i2}]$

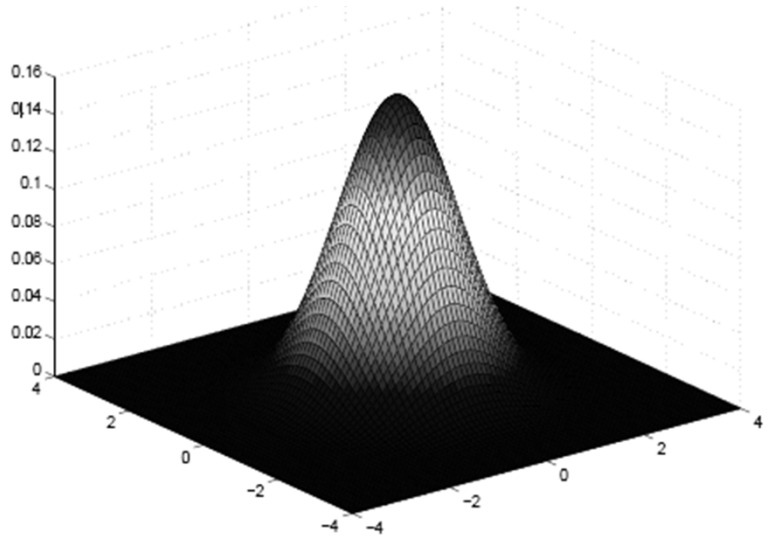
$$f(x_1, x_2) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right]$$

Where

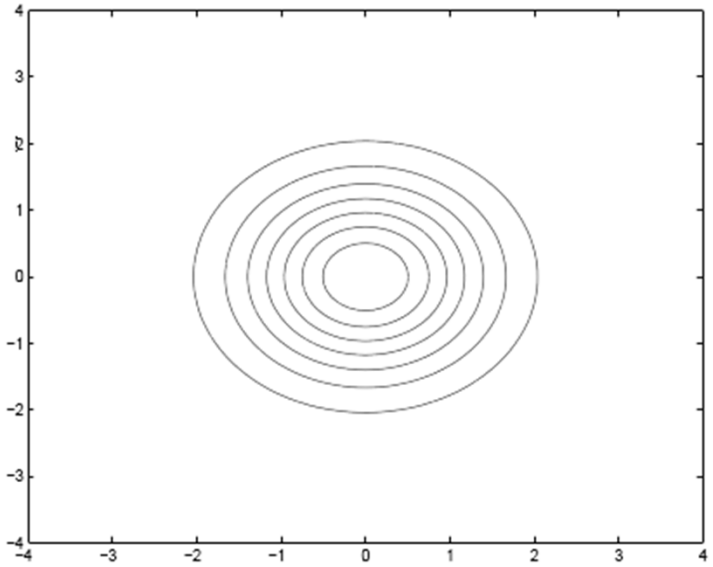
$$z = \frac{\left(x_{i1} - \mu_{x_1}\right)^2}{\sigma_{x_1}^2} - \frac{2\rho\left(x_{i1} - \mu_{x_1}\right)\left(x_{i2} - \mu_{x_2}\right)}{\sigma_1\sigma_2} + \frac{\left(x_{i2} - \mu_{x_2}\right)^2}{\sigma_{x_2}^2}$$

Bivariate Normal Plot #1

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



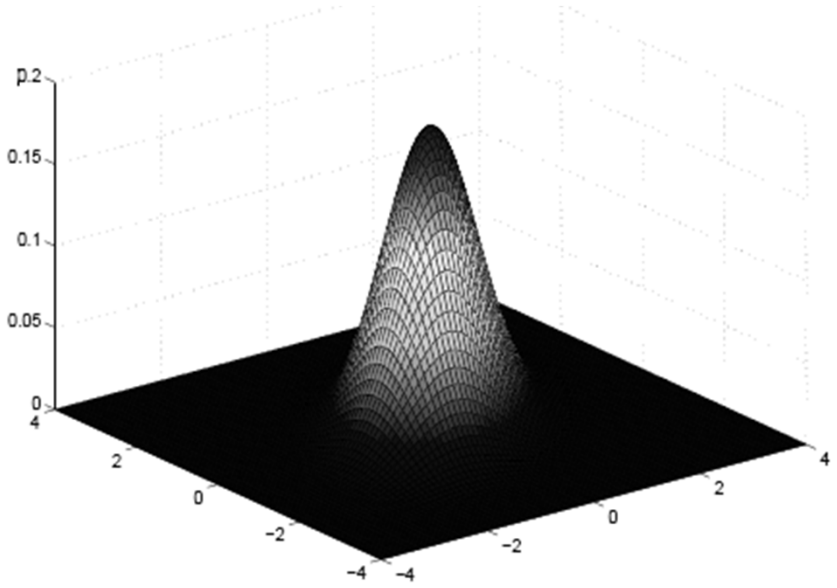
Density Surface (3D)



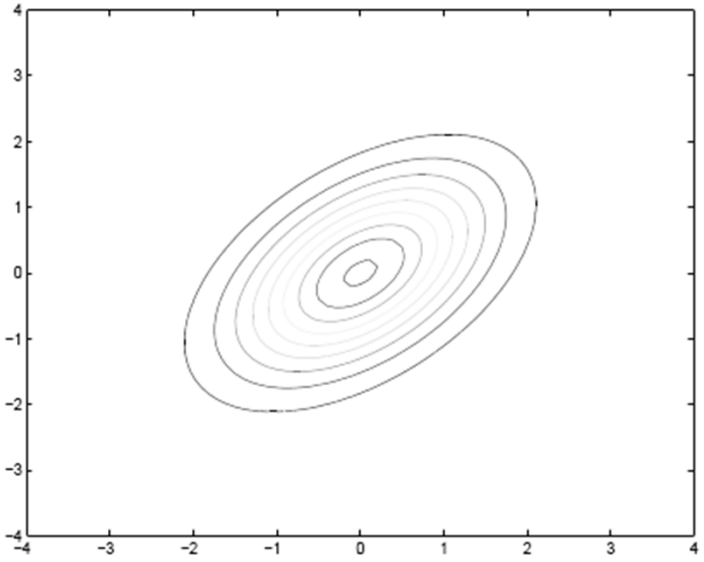
Density Surface (2D):
Contour Plot

Bivariate Normal Plot #2

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{x_1} & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma^2_{x_2} \end{bmatrix} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$$



Density Surface (3D)



Density Surface (2D):
Contour Plot

MULTIVARIATE DISTRIBUTIONS (VARIABLES ≥ 2)

Multivariate Normal Distribution

- The multivariate normal distribution is the generalization of the univariate normal distribution to multiple variables
  - The bivariate normal distribution just shown is part of the MVN

- The MVN provides the relative likelihood of observing **all** *p* variables for a subject *i* simultaneously:

$$\mathbf{x}_i = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip}]$$

- The multivariate normal density function is:

$$f(\mathbf{x}_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2}\right]$$

The Multivariate Normal Distribution

$$f(\mathbf{x}_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2}\right]$$

- The mean vector is $\boldsymbol{\mu} = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \vdots \\ \mu_{x_p} \end{bmatrix}$

- The covariance matrix is $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_p} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \cdots & \sigma_{x_2x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_1x_p} & \sigma_{x_2x_p} & \cdots & \sigma_{x_p}^2 \end{bmatrix}$

  - The covariance matrix must be non-singular (invertable)

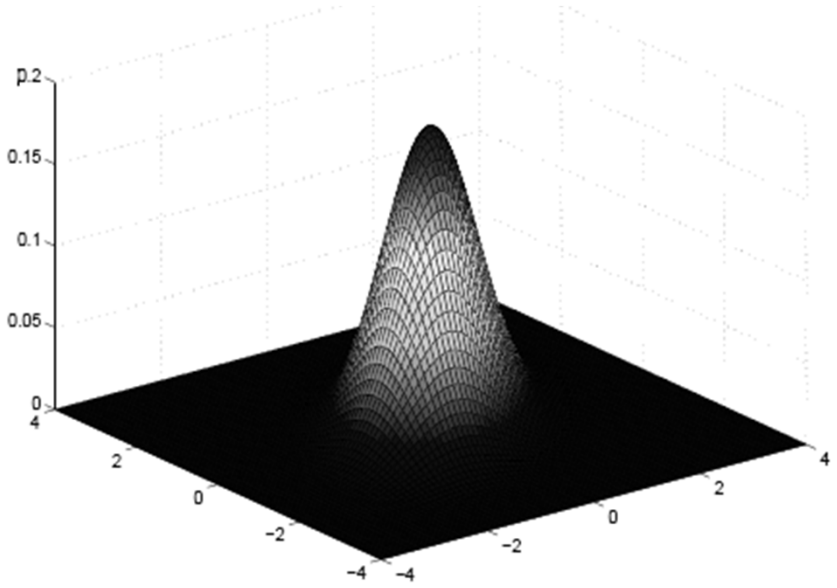
Multivariate Normal Notation

- Standard notation for the multivariate normal distribution of *p* variables is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Our bivariate normal would have been $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

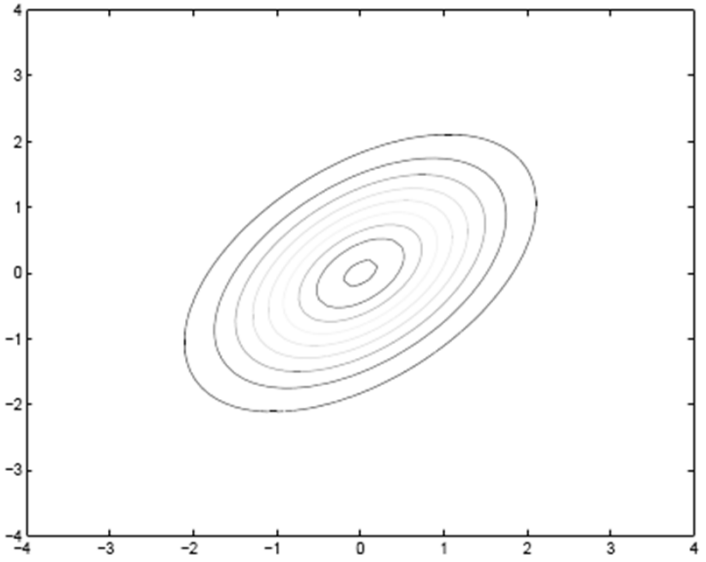
- **<u>In data:</u>**
  - The multivariate normal distribution serves as the basis for most every statistical technique commonly used in the social and educational sciences
    - General linear models (ANOVA, regression, MANOVA)
    - General linear mixed models (HLM/multilevel models)
    - Factor and structural equation models (EFA, CFA, SEM, path models)
    - Multiple imputation for missing data
  - Simply put, the world of commonly used statistics revolves around the multivariate normal distribution
    - Understanding it is the key to understanding many statistical methods

Bivariate Normal Plot #2 (Multivariate Normal)

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$$



Density Surface (3D)



Density Surface (2D):
Contour Plot

Multivariate Normal Properties

- The multivariate normal distribution has some useful properties that show up in statistical methods

- If $\mathbf{X}$ is distributed multivariate normally:

1. Linear combinations of $\mathbf{X}$ are normally distributed
2. All subsets of $\mathbf{X}$ are multivariate normally distributed
3. A zero covariance between a pair of variables of $\mathbf{X}$ implies that the variables are independent
4. Conditional distributions of $\mathbf{X}$ are multivariate normal

Sampling Distributions of MVN Statistics

- Just like in univariate statistics, there is a multivariate central limit theorem

If the set of N observations on p variables is multivariate normal or not:

- The distribution of the mean vector is:

$$\bar{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu}_{\mathbf{X}}, \frac{\boldsymbol{\Sigma}_{\mathbf{X}}}{N}\right)$$

- The distribution of $(N-1)\mathbf{S}_{\mathbf{X}}$ (covariance matrix) is Wishart (a multivariate chi-square) with degrees of freedom $N-1$

$$W_p(N-1, \boldsymbol{\Sigma}_{\mathbf{X}})$$

The Wishart Distribution

- The Wishart distribution is a multivariate chi-square distribution

$$W_{N-1}(\mathbf{S},\boldsymbol{\Sigma}) = \frac{|\mathbf{S}|^{\frac{(N-p-2)}{2}} \exp\left[\frac{tr\mathbf{S}\boldsymbol{\Sigma}^{-1}}{2}\right]}{2^{\left(\frac{p(N-1)}{2}\right)}\pi^{\frac{p(p-1)}{4}}|\boldsymbol{\Sigma}|^{\frac{N-1}{2}}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(N-i)\right)}$$

  - Input: **S** (model predicted covariance matrix)
    …Output: Likelihood value
  - Fixed: $\boldsymbol{\Sigma}$ (sample value of covariance matrix)

- In statistics and SEM, it appears whenever:
  - Data are assumed multivariate normal
  - Only covariance matrix-based inferences are needed
    - Mean vector ignored
  - Mainly: Initial ML factor analysis and structural equation modeling

Sufficient Statistics

- The sample estimates $\bar{\mathbf{x}}$ and $\mathbf{S}$ are called sufficient statistics
  - All of the information contained in the data can be summarized by these two statistics alone
    - No data is needed for analysis – only these statistics

- Only true when:
  - Data truly follow a multivariate normal distribution
  - No missing data are present

INTRODUCTION TO MPLUS

The Mplus Statistical Package

- Mplus provides a general latent variable modeling framework that allows for combinations of:
  - Continuous or categorical latent variables
  - Continuous, categorical, count, nominal or censored data

- Mplus is commercial software that is available through the Holland Computing Center
  - See document on course main page for how to:
    - Create a an account
    - Run Mplus

- Mplus is also available for purchase:
  - Available at http://www.statmodel.com
  - From $195 to $350 (student)

Mplus Data File Input Format

- Mplus input files must be ASCII text based (so not binary)
  - Text-based file formats: *.txt, *.dat, *.csv
  - Not-text-based file formats: *.xlsx, *.sas7bdat, *.sav

- The easiest way to get data files into Mplus is to use "free-formatting" (some type of delimiter between columns)
  - I prefer comma-delimited files and will only use those in this course
  - Cannot start with variable names in first row of data

- Typically, I store data in Excel and save as a comma-delimited file
  - Save As…*.csv…
    - (then click OK to the first question)…(then click YES to the second)
    - Ignore the warning (click NO) to re-save when closing the Excel Workbook

Mplus Syntax Conventions

- Most syntax must have a semi-colon end each line (;)
  - Exceptions: TITLE section, comments, and continuing lines
- Comments are denoted with an exclamation point (!)
- Syntax is organized by sections; headings of sections end with colons (:)
  - TITLE:, DATA:, VARIABLE:, DEFINE:, and MODEL: are what we use this week
- Syntax cannot exceed 90 characters per row (ꟾ)
- Mplus input files are typically saved with the extension *.inp
- Mplus output files are typically saved with the extension *.out
  - Both are ASCII text (i.e., you can open with text editors)
- The default location for the data file and output file are the folder containing the input file

Mplus TITLE Section

- The TITLE section contains the label of the analysis
- You can type whatever you want here…it will appear verbatim at the top of the output file
- You do not have to terminate this section with a semi-colon
- This section is optional

Mplus DATA Section

- The DATA section is where data files are defined

- Define the name (and path if different from input file folder) by using the command:

FILE = mydata.csv

- POTENTIAL MISTAKES:
  - Data must be numeric – if not errors happen
  - First row of data should not contain variable names

- This section is NOT OPTIONAL

Mplus VARIABLE Section

- The Mplus VARIABLE section defines the names of the variables in the data file, variable types, and variables in your analysis

  - NAMES = provides variable names
    - Names cannot be more than 8 characters
    - Lists of variables can be created (i.e., X1-X10 makes 10 variables)
    - By default all variables listed in the NAMES section are assumed to be part of the analysis
  - USEVARIABLE = provides names of variables used in the analysis (optional)
  - IDVARIABLE = provides the ID variable name (optional)

- This section is NOT OPTIONAL

Mplus DEFINE Section

- The DEFINE section is where new variables are created

- To test our equal slopes hypothesis we created interaction variables by the following syntax:

```
TITLE: !title section puts text below at the top of the output file
    ANCOVA MODEL WITH GENDER, MOTIVATION, AND ACHIEVEMENT
    TESTING INTERACTIONS - DIFFERENT SLOPES WITHIN GENDER GROUP
DATA: !data section defines data file
    FILE = exampledata.csv;

VARIABLE: !variable section defines varibles in data file
    NAMES = ID achieve selfest motivate gender se1-se5
            motiv1-motiv5 ach1-ach20;
    IDVARIABLE = ID;
    USEVARIABLE = achieve selfest gender motivate femaleSE femaleM;

DEFINE:
    femaleSE = gender*selfest;
    femaleM = gender*motivate;

MODEL:
    achieve ON selfest gender motivate femaleSE femaleM;
```

Mplus MODEL Section

- The MODEL section is where you define the model

- The only models we ran this week were GLMs
  - These models use the ON statement (ON = REGRESSION)

    achieve ON selfest motivate

- And an empty model to figure out the covariance matrix of the MOTIVATION items
  - This used the WITH statement (WITH = COVARIANCE)

    motiv1-motive5 WITH motive1-motive5

WRAPPING UP

Wrapping Up

- Today was an introduction to…
  - Our course structure
  - Mplus

- …and a review of:
  - Topics in matrix algebra
  - Multivariate normal distributions

- All of these topics will follow us throughout the semester

- Homework #1: available online – due next Wednesday at 11:59am (before class); submitted via email only