Univariate and Multivariate Statistical Distributions

Multivariate Methods in Education
ERSH 8350
Lecture #3 – August 31, 2011
Today’s Lecture

• Univariate and multivariate statistics

• The univariate normal distribution

• The multivariate normal distribution

• How to inspect whether data are normally distributed
Motivation for Today’s Lecture

• The univariate and multivariate normal distributions serve as the backbone of the most frequently used statistical procedures
  ➢ They are very robust and very useful

• Understanding their form and function will help you learn about a whole host of statistical routines immediately
  ➢ This course is essentially a course based in normal distributions
Today’s Example

• For today, we will use hypothetical data on the heights and weights of 50 students picked at random from the University of Georgia.

• You can follow along in SAS using the example syntax and the data file “heightweight.csv”.

![Graph showing scatter plot of height vs weight]
Random Variables

A random variable is a variable whose outcome depends on the result of chance

- Can be continuous or discrete
  - Example of continuous: $x$ represents the height of a person, drawn at random
  - Example of discrete: $x$ represents the gender of a person, drawn at random
- Had a density function $f(x)$ that indicates relative frequency of occurrence
  - A density function is a mathematical function that gives a rough picture of the distribution from which a random variable is drawn

We will discuss the ways that are used to summarize a random variable $x$
UNIVARIATE STATISTICS
Mean

The **mean** is a measure of central tendency

- For a single variable, $\mathbf{x}$ (put into a column vector):

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} \mathbf{x}^T \mathbf{1}$$

($\mathbf{1}$ is a column vector with elements all 1 – size $N \times 1$)

- The sampling distribution of $\bar{x} \sim N \left( \mu_x, \frac{\sigma_x^2}{N} \right)$
  - Converges to normal as $N$ increases (central limit theorem)

- In our data:
  - mean height: $\bar{x}_1 = 67.28$ (in inches)
**Variance**

The **variance** is a measure of variability (or spread/range)

- For a single variable, $\mathbf{x}$ (put into a column vector):
  \[
  s_x^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \frac{1}{N} (\mathbf{x} - \bar{x} \mathbf{1})^T (\mathbf{x} - \bar{x} \mathbf{1})
  \]

- Note, if divided by $N$, maximum likelihood likelihood estimate is used
  - MLE is biased, so often $N-1$ is used

- If the original data are normally distributed $x \sim N(\mu_x, \sigma_x^2)$ then the sampling distribution of the variance is:
  \[
  (N - 1) \frac{s_x^2}{\sigma_x^2} \sim \chi_{N-1}^2
  \]

- For our data:
  - Variance of height: $s_1^2 = 21.8$ (in inches squared)
  - Standard deviation of height: $s_1 = 4.7$ (in inches)
Linear Combinations of $x$

If we create a new variable based on a linear combination of our previous variable: $y = b + ax$

- The mean of the linear combination is
  $\bar{y} = b + a\bar{x}$

- The variance of the linear combination is
  $s_y^2 = a^2 s_x^2$

- In our data:
  - If we wanted to change height from inches to centimeters, we would make the linear combination $y_{cm} = 2.54x_{in}$
  - Mean of height in CM: $\bar{y} = 2.54\bar{x} = 170.9$
  - Variance of height in CM$^2$: $s_y^2 = 2.54^2 s_x^2 = 140.4$
  - Standard deviation of height in CM: $s_y = 2.54 s_x = 11.8$
UNIVARIATE STATISTICAL DISTRIBUTIONS
Univariate Normal Distribution

- For a continuous random variable \( x \) (ranging from \(-\infty\) to \(\infty\)) 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 \( \mu_x \)
  - The variance \( \sigma_x^2 \)

- 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)
Chi-Square Distribution

- Another frequently used univariate distribution is the Chi-Square distribution
  - Previously we mentioned the sampling distribution of the variance followed a Chi-Square distribution

- For a continuous (i.e., quantitative) random variable $x$ (ranging from 0 to $\infty$), the chi-square distribution is given by:

$$ f(x) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{\nu/2 - 1} \exp \left( -\frac{x}{2} \right) $$

- $\Gamma(\cdot)$ is called the gamma function
- The chi-square distribution is governed by one parameter: $\nu$ (the degrees of freedom)
  - The mean is equal to $\nu$; the variance is equal to $2\nu$
(Univariate) Chi-Square Distribution

$$f(x)$$

$$x$$
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
Assessing Distributional Assumptions Graphically

- A useful tool to evaluate the plausibility of a distributional assumption is that of the Quantile versus Quantile Plot (Q-Q plot).

- A Q-Q plot is formed by comparing the observed quantiles of a variable with that of a known statistical distribution.
  - A quantiles is the particular ordering of a given observation.
  - In our data, a person with a height of 71 is the 39th tallest person (out of 50).
  - This would correspond to the person being at the $\frac{39-.5}{50} = .77$ or .77 quantile of the distribution.
    - Taller than 77% of the distribution, or at the 77th percentile.
Q-Q Plot

- A Q-Q Plot is built by:
  1. Ordering the data from smallest to largest
  2. Calculating the quantiles of each data point
  3. Looking up the quantiles of each data point using a known statistical distribution
  4. Plotting the data against the value predicted by the theoretical statistical distribution

- If the data deviate from a straight line, the data are not likely to follow from that theoretical distribution
Q-Q Plot Example

- Using the example Excel spreadsheet, Q-Q plots of our height data were built, comparing the data with a normal distribution (left) and a chi-square distribution (right).
  - The line is where the data “should” be if they followed that distribution.
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 it independent of that of weight)
  ➢ Could have looked at weight, marginally

• Multivariate statistics is about exploring joint distributions
  ➢ How 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: \(X\) (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

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

- Here \( \mathbf{1} \) is a \( N \times 1 \) vector of 1s
- The resulting mean vector is a \( p \times 1 \) vector of means

- For our data:

\[
\bar{x} = \begin{bmatrix}
67.2 \\
154.5
\end{bmatrix} = \begin{bmatrix}
\bar{x}_{height} \\
\bar{x}_{weight}
\end{bmatrix}
\]
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

\[
s_{x_1x_2} = \frac{1}{N} \sum_{i=1}^{N} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)
\]

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

- 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:
\[
\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1} = \begin{bmatrix}
\sqrt{S_{x_1}^2} & \cdots & \sqrt{S_{x_1x_p}} \\
\sqrt{S_{x_1}^2} & \ddots & \sqrt{S_{x_p}^2} \\
\sqrt{S_{x_1x_p}} & \cdots & \sqrt{S_{x_p}^2}
\end{bmatrix} = \begin{bmatrix}
1 & \cdots & r_{x_1x_p} \\
\vdots & \ddots & \vdots \\
r_{x_1x_p} & \cdots & 1
\end{bmatrix}
\]
Example Covariance Matrix

- For our data, the covariance matrix was:
  \[ S = \begin{bmatrix} 21.8 & 158.6 \\ 158.6 & 2,260.2 \end{bmatrix} \]

- The diagonal matrix \( D \) was:
  \[ D = \begin{bmatrix} \sqrt{21.8} & 0 \\ 0 & \sqrt{2,260.2} \end{bmatrix} \]

- The correlation matrix \( R \) was:
  \[ R = D^{-1}SD^{-1} = \begin{bmatrix} \frac{1}{\sqrt{21.8}} & 0 \\ 0 & \frac{1}{\sqrt{2,260.2}} \end{bmatrix} \begin{bmatrix} 21.8 & 158.6 \\ 158.6 & 2,260.2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{21.8}} & 0 \\ 0 & \frac{1}{\sqrt{2,260.2}} \end{bmatrix} \]
  \[ R = \begin{bmatrix} 1.00 & 0.72 \\ 0.72 & 1.00 \end{bmatrix} \]
Generalized Variance

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

  \[ \text{Generalized Sample Variance} = |S| \]

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

• **In data:**
  - 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

\[
Total \ Sample \ Variance = \sum_{i=1}^{p} s_{x_i}^2 = \text{tr} \ S
\]

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

- In data:
  - 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, \(\mathbf{x}_i = [x_{i1} \ x_{i2}]\)

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

Where

\[
z = \frac{(x_{i1} - \mu_{x_1})^2}{\sigma_{x_1}^2} - \frac{2\rho(x_{i1} - \mu_{x_1})(x_{i2} - \mu_{x_2})}{\sigma_1\sigma_2} + \frac{(x_{i2} - \mu_{x_2})^2}{\sigma_{x_2}^2}
\]
Bivariate Normal Plot #1

\[ \mu = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]
Bivariate Normal Plot #2

$$\mu = \begin{bmatrix} \mu x_1 \\ \mu x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$$
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:
  \[ x_i = [x_{i1} \ x_{i2} \ \ldots \ x_{ip}] \]

- The multivariate normal density function is:
  \[
  f(x_i) = \frac{1}{p \pi^{\frac{p}{2}} |\Sigma|^\frac{1}{2}} \exp \left[ -\frac{(x_i^T - \mu)^T \Sigma^{-1} (x_i^T - \mu)}{2} \right]
  \]
The Multivariate Normal Distribution

\[ f(x_i) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x_i^T - \mu)^T \Sigma^{-1} (x_i^T - \mu) \right] \]

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

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

- The covariance matrix must be non-singular (invertable)
Multivariate Normal Notation

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

• **In data:**
  - 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)

\[
\mu = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}
\]
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
Property #1: Linear Combinations

- If \( \mathbf{X} \sim N_p(\mu, \Sigma) \) then any set of \( q \) linear combinations of variables \( \mathbf{A}_{(q \times p)} \) are also normally distributed as:
  \[
  \mathbf{X}\mathbf{A}^T \sim N_q(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T)
  \]

- If we had three variables \( (p = 3) \) and we wanted to create a new variable \( Y \) that was the difference between \( X_1 \) and \( X_2 \) then:

  \[
  \mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}
  \]

  For \( \mathbf{X} \):

  \[
  \begin{align*}
  \mu_X &= \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \mu_{x_3} \end{bmatrix}, & \Sigma_X &= \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \sigma_{x_1x_3} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \sigma_{x_2x_3} \\ \sigma_{x_1x_3} & \sigma_{x_2x_3} & \sigma_{x_3}^2 \end{bmatrix}
  \end{align*}
  \]

  For \( \mathbf{Y} = \mathbf{X}\mathbf{A}^T \):

  \[
  \begin{align*}
  \mu_Y &= \begin{bmatrix} \mu_{x_1} - \mu_{x_2} \end{bmatrix}, & \Sigma_Y &= \begin{bmatrix} \sigma_{x_1}^2 + \sigma_{x_2}^2 - 2\sigma_{x_1x_2} \end{bmatrix}
  \end{align*}
  \]
The linear combinations property is useful in several areas of statistics:

- Principal components (can determine mean and variance of each component)
- Regression/ANOVA (can determine error of prediction for Y)

We will see this property later in the course
Property #2: All Subsets are MVN

- If we had three variables \( (p = 3) \) and we wanted to know the distribution of \( X_1 \) and \( X_2 \) (\( \mathbf{X}^* = [X_1 \quad X_2] \)) then:

\[
\mathbf{\mu}_X = \begin{bmatrix}
\mu_{x_1} \\
\mu_{x_2} \\
\mu_{x_3}
\end{bmatrix} = \begin{bmatrix}
\mathbf{\mu}_{X^*}
\end{bmatrix}
\]

\[
\Sigma_X = \begin{bmatrix}
\sigma_{x_1}^2 & \sigma_{x_1x_2} & \sigma_{x_1x_3} \\
\sigma_{x_2} & \sigma_{x_2}^2 & \sigma_{x_2x_3} \\
\sigma_{x_1x_3} & \sigma_{x_2x_3} & \sigma_{x_3}
\end{bmatrix} = \begin{bmatrix}
\Sigma_{X^*} & \Sigma_{X^*x_3} \\
\Sigma_{x_3x^*} & \Sigma_{x_3}
\end{bmatrix}
\]

- Then \( \mathbf{X}^* \sim N_2(\mathbf{\mu}_{X^*}, \Sigma_{X^*}) \)
- **In data:** all univariate distributions must be normal
Property #3: Zero Covariance = Independence

• As we will see shortly, all of the information about the multivariate normal distribution is contained within the mean vector and the covariance matrix
  ➢ No other statistics are needed if you have these

• Therefore, if a pair of variables has a zero covariance (meaning zero correlation), then they are independent
  ➢ No other way dependency can happen (no higher moments that are variables)

• Important: this property is not true of all distributions
  ➢ Especially distributions for categorical data
Property #4: Conditional Distributions of Variables are Multivariate Normal

- The conditional distribution of sets of variables from a MVN is also MVN
  - This is how missing data get imputed
- If we were interested in the distribution of the first $q$ variables, we partition three matrices:
  - The data: $\begin{bmatrix} \mathbf{X}_1:(N \times q) & \mathbf{X}_2:(N \times p-q) \end{bmatrix}$
  - The mean vector: $\begin{bmatrix} \mathbf{\mu}_1:(q \times 1) \\ \mathbf{\mu}_2:(p-q \times 1) \end{bmatrix}$
  - The covariance matrix: $\begin{bmatrix} \mathbf{\Sigma}_{11}:(q \times q) & \mathbf{\Sigma}_{12}:(q \times p-q) \\ \mathbf{\Sigma}_{21}:(p-q \times q) & \mathbf{\Sigma}_{22}:(p-q \times p-q) \end{bmatrix}$
Conditional Distributions of MVN Variables

- The conditional distribution of $X_1$ given the values of $X_2 = x_2$ is then:

$$X_1 | X_2 \sim N_q(\mu^*, \Sigma^*)$$

Where (using our partitioned matrices):

$$\mu^* = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2' - \mu_2)$$

And:

$$\Sigma^* = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$
Property #4: Conditional Distributions of Variables are Multivariate Normal

- If we had three variables \( (p = 3) \) and we wanted to know the distribution of \( X_1 \) and \( X_2 \) \( (X^* = [X_1 \quad X_2]) \) conditional on the value of \( X_3 \) then:

\[
\mu_X = \begin{bmatrix}
\mu_{x_1} \\
\mu_{x_2} \\
\mu_{x_3}
\end{bmatrix} = \begin{bmatrix}
\mu_{x^*} \\
\mu_{x_3}
\end{bmatrix}
\]

\[
\Sigma_X = \begin{bmatrix}
\sigma_{x_1}^2 & \sigma_{x_1 x_2} & \sigma_{x_1 x_3} \\
\sigma_{x_1 x_2} & \sigma_{x_2}^2 & \sigma_{x_2 x_3} \\
\sigma_{x_1 x_3} & \sigma_{x_2 x_3} & \sigma_{x_3}^2
\end{bmatrix} = \begin{bmatrix}
\Sigma_{x^*} & \Sigma_{x^* x_3} \\
\Sigma_{x_3 x^*} & \Sigma_{x_3}
\end{bmatrix}
\]

\[
X^*|X_3 \sim N_2\left(\mu_{x^*} + \Sigma_{x^* x_3} \Sigma_{x_3}^{-1}(x'_3 - \mu_3), \Sigma_{x^*} - \Sigma_{x^* x_3} \Sigma_{x_3}^{-1} \Sigma_{x_3 x^*}\right)
\]
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:

1. The distribution of the mean vector is:

\[
\bar{\mathbf{x}} \sim \mathcal{N}_p \left( \mu_x, \frac{\Sigma_x}{N} \right)
\]

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

\[
\mathcal{W}_p (N - 1, \Sigma_x)
\]
The Wishart Distribution

The Wishart distribution is a multivariate chi-square distribution

\[ W_{N-1}(S, \Sigma) = \frac{1}{2^{p(N-1)/2} \pi^{p(p-1)/4} |\Sigma|^{(N-1)/2}} \exp \left(\frac{trSS^{-1}}{2}\right) \]

- Input: \( S \) (model predicted covariance matrix)
  - Output: Likelihood value
- Fixed: \( \Sigma \) (sample value of covariance matrix)

In statistics, 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 $\bar{x}$ and $S$ 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
ASSESSING MULTIVARIATE NORMALITY
Assessing Multivariate Normality of Data

- A large number of statistical methods rely upon the assumption data follow a multivariate normal distribution
  - Data “should” be multivariate normal then

- Methods exist for getting a rough picture of whether or not data are multivariate normal
  - I will highlight a graphical method using the Q-Q plot

- In reality:
  - Determining exact distribution of data is difficult if not impossible
  - Methods that assume multivariate normality are fairly robust to small deviations from that assumption
Re-examining the Multivariate Normal Distribution

- The multivariate normal distribution density is:

\[
f(x_i) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left[-\frac{(x_i^T - \mu)^T \Sigma^{-1} (x_i^T - \mu)}{2}\right]
\]

- The term in the exponent \((x_i^T - \mu)^T \Sigma^{-1} (x_i^T - \mu)\) gives a scalar number that represents the squared “statistical” distance of a person’s observed variables to the mean
  - Mahalanobis distance squared
  - Distance contribution for each variable is standardized

- \((x_i^T - \mu)^T \Sigma^{-1} (x_i^T - \mu)\) is distributed \(\chi^2_p\)
  - We now have a way we can construct a Q-Q plot for assessing multivariate normality
Q-Q Plot for Multivariate Normality

- A Q-Q Plot for multivariate normality is built by:
  1. Calculating \( d_i = (x_i^T - \bar{x})^T S^{-1}(x_i^T - \bar{x}) \) using the sample estimated mean vector and covariance matrix
  2. Ordering the \( d_i \) from smallest to largest across all observations
  3. Calculating the quantiles of each data point
  4. Looking up the quantiles of each data point using a \( \chi^2_p \) distribution (df = # of variables)
  5. Plotting the data against the value predicted by the theoretical statistical distribution
Good News: SAS Macro for Assessing Multivariate Normality

- Although the process of assessing multivariate normality sounds tedious, our friends at SAS have simplified things considerably
  - MULTNORM Macro available at: http://support.sas.com/kb/24/983.html

- The MULTNORM macro evaluates both univariate and multivariate using a set of hypothesis tests and plots
  - The hypothesis tests can be difficult to use: very sensitive to violations of normality
  - The Q-Q plots are better
    - Big question: how much deviation is too much?
    - Typical answer (from practice): doesn’t matter (no one checks)
Assessing the Multivariate Normality of Teacher Rating Data

- To demonstrate how to assess multivariate normality we will use the 14 item teacher ratings data

- For multivariate normality to hold data must be:
  - Univariate normal for all items individually (marginally)
  - Multivariate normal for all item jointly

- Steps in the process:
  - 1. Check each variable by itself using Q-Q plots and Shapiro-Wilk Hypothesis tests
    - Do all variables individually follow a univariate normal distribution?
      - YES: Go to step #2
      - NO: Stop here
  - 2. Check all variables simultaneously using Q-Q plots and Mardia hypothesis tests
    - Do all variables together follow a multivariate normal distribution?
      - YES: Congrats...you are in a rare state
      - NO: Knowledge is power
Hypothesis Tests

**Univariate Tests**

**Shapiro-Wilk univariate hypothesis test:**
- $H_0$: Data are univariate normal
- $H_A$: Data are not univariate normal

**Based on these results:** since univariate normality does not hold, multivariate normality cannot hold (we will continue on anyway)

**Mardia/Henze-Zirkler Multivariate hypothesis test:**
- $H_0$: Data are multivariate normal
- $H_A$: Data are not multivariate normal
Problems with Hypothesis Tests

• Shapiro-Wilk test is very powerful and may lead to rejection of univariate normality for larger sample sizes

• Multivariate tests also may have issues in that p-values would lead to rejecting multivariate normality

• Better approach: look at plots and be sure “large” deviations are not present
  ➢ “large” is an open question
Univariate Plots

MULTNORM macro: Univariate and Multivariate Normality Tests

Item 12

MULTNORM macro: Univariate and Multivariate Normality Tests

Item 13

MULTNORM macro: Univariate and Multivariate Normality Tests

Item 14

MULTNORM macro: Univariate and Multivariate Normality Tests

Item 15
Multivariate Q-Q Plot of Squared Mahalanobis Distance

MULTNORM macro: Chi-square Q-Q plot
Checking Normality Again: Height and Weight Data

- Let’s now check for multivariate normality of our height and weight data from the beginning of class:

```
MULTNORM macro: Univariate and Multivariate Normality Tests

The MODEL Procedure

<table>
<thead>
<tr>
<th>Equation</th>
<th>Normality Test</th>
<th>Test Statistic</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>Shapiro-Wilk W</td>
<td>0.97</td>
<td>0.4895</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>Shapiro-Wilk W</td>
<td>0.99</td>
<td>0.9534</td>
<td></td>
</tr>
<tr>
<td>System</td>
<td>Mardia Skewness</td>
<td>1.81</td>
<td>0.7709</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mardia Kurtosis</td>
<td>-0.86</td>
<td>0.3916</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Henze-Zirkler T</td>
<td>-2.77</td>
<td>0.0056</td>
<td></td>
</tr>
</tbody>
</table>
```

- Univariate tests:
  - P-values are large :: univariate normality likely for both variables

- Multivariate tests:
  - P-values large for Mardia tests, small for Henze-Zirkler

- Conclusion from tests: likely multivariate normal
Graphical Inspection
When and Why to Check Normality

• Good news: most methods are fairly robust to violations of normality
  ➢ For homework we will have to check this assumption

• Methods taught in class where normality should be checked:
  ➢ Missing data imputation (unconditional data)
  ➢ Linear mixed models (conditional data - residuals)
  ➢ Exploratory Factor Analysis (unconditional data)
  ➢ Confirmatory Factor Analysis (unconditional data)
  ➢ Structural Equation Modeling (unconditional and conditional data)
  ➢ Mixture models assuming normality (conditional data)

• If normality does not hold:
  ➢ Check your data for input errors (i.e., 99 = missing)
  ➢ Plot every variable – visualization is the most important thing
  ➢ Do not “clean” your data
  ➢ Use a model with a different link function, if possible (generalized model)
WRAPPING UP
Concluding Remarks

• The univariate and multivariate normal distributions serve as the backbone of the most frequently used statistical procedures
  ➢ They are very robust and very useful

• Understanding their form and function will help you learn about a whole host of statistical routines immediately
  ➢ This course is essentially a course based in normal distributions

• Inspecting normality is a tricky thing
  ➢ More often than not, data will not be normal
  ➢ More often than not, no one will check (don’t ask, don’t tell)

• Next week: linear models in matrices, repeated measures ANOVA, Multivariate ANOVA
  ➢ Classical statistics – described as a background for modern methods