Multivariate Normal Distribution

Lecture 4

July 21, 2011

Advanced Multivariate Statistical Methods

ICPSR Summer Session #2

Last Time

- Matrices and vectors
  - Eigenvalues
  - Eigenvectors
  - Determinants
- Basic descriptive statistics using matrices:
  - Mean vectors
  - Covariance Matrices
  - Correlation Matrices

Today's Lecture

- Putting our new knowledge to use with a useful statistical distribution: the Multivariate Normal Distribution
- This roughly maps onto Chapter 4 of Johnson and Wichern

Multivariate Normal Distribution

- The generalization of the univariate normal distribution to multiple variables is called the multivariate normal distribution (MVN)

- Many multivariate techniques rely on this distribution in some manner

- Although real data may never come from a true MVN, the MVN provides a robust approximation, and has many nice mathematical properties

- Furthermore, because of the central limit theorem, many multivariate statistics converge to the MVN distribution as the sample size increases

Univariate Normal Distribution

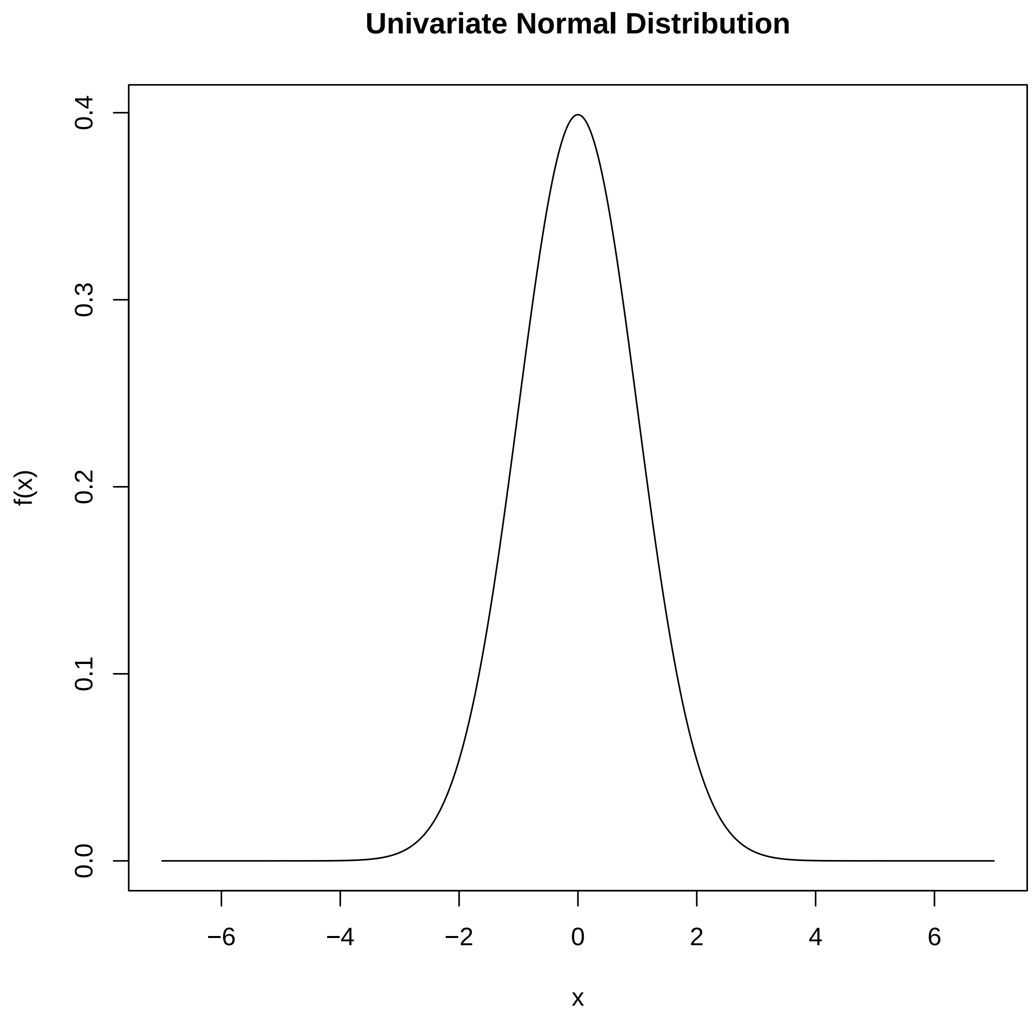
- The univariate normal distribution function is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2}$$

- The mean is μ
- The variance is σ^2
- The standard deviation is σ
- Standard notation for normal distributions is $N(\mu, \sigma^2)$, which will be extended for the MVN distribution

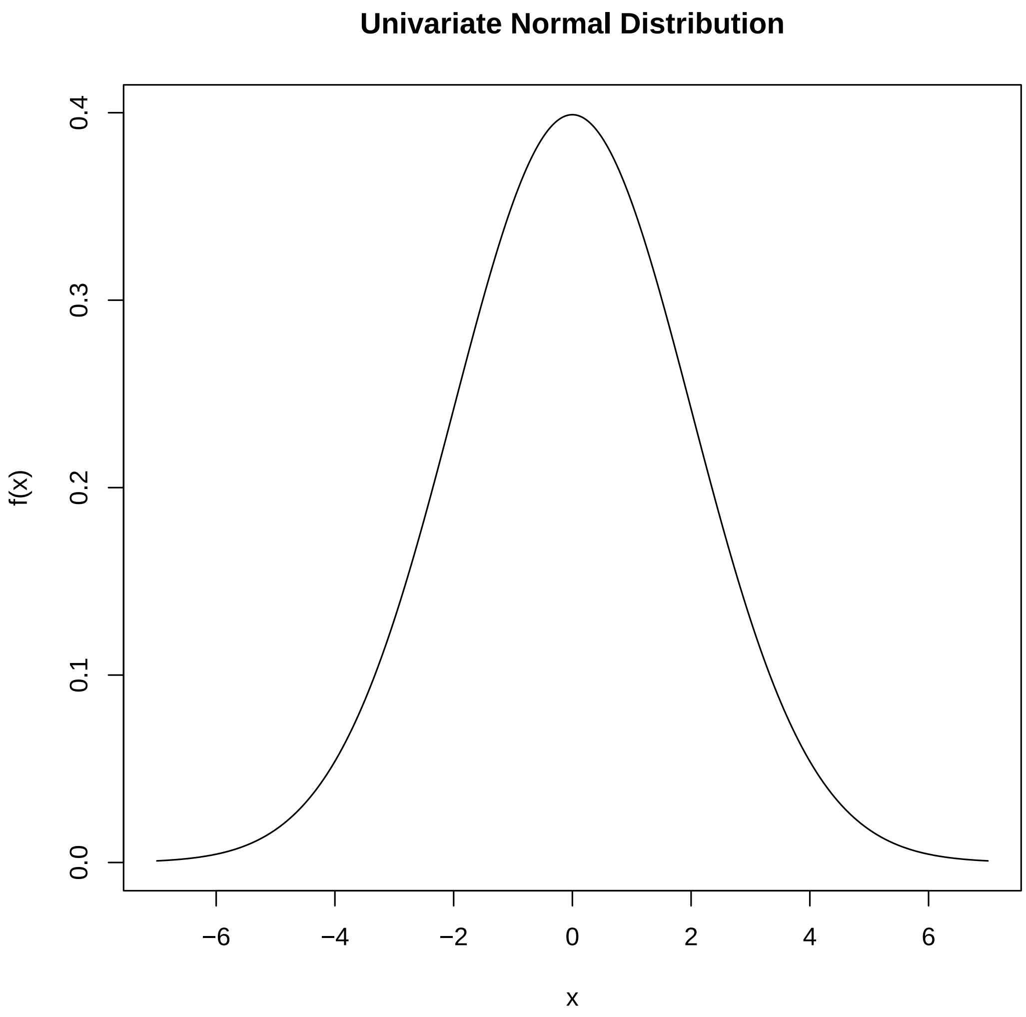
Univariate Normal Distribution

$$N(0,1)$$



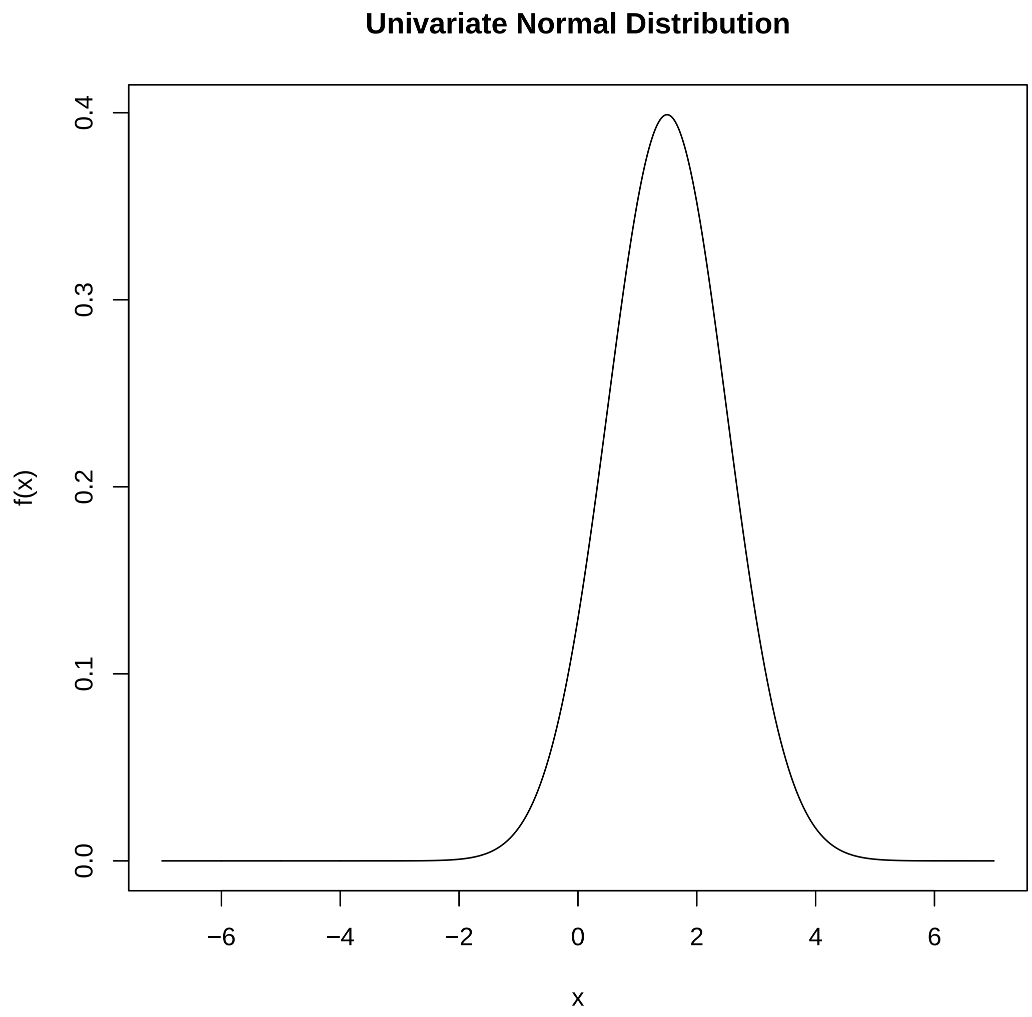
Univariate Normal Distribution

$$N(0,2)$$



Univariate Normal Distribution

$$N(1.75, 1)$$



UVN - Notes

- The area under the curve for the univariate normal distribution is a function of the variance/standard deviation

- In particular:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.954$$

- Also note the term in the exponent:

$$\left(\frac{(x-\mu)}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

- This is the square of the distance from x to μ in standard deviation units, and will be generalized for the MVN

MVN

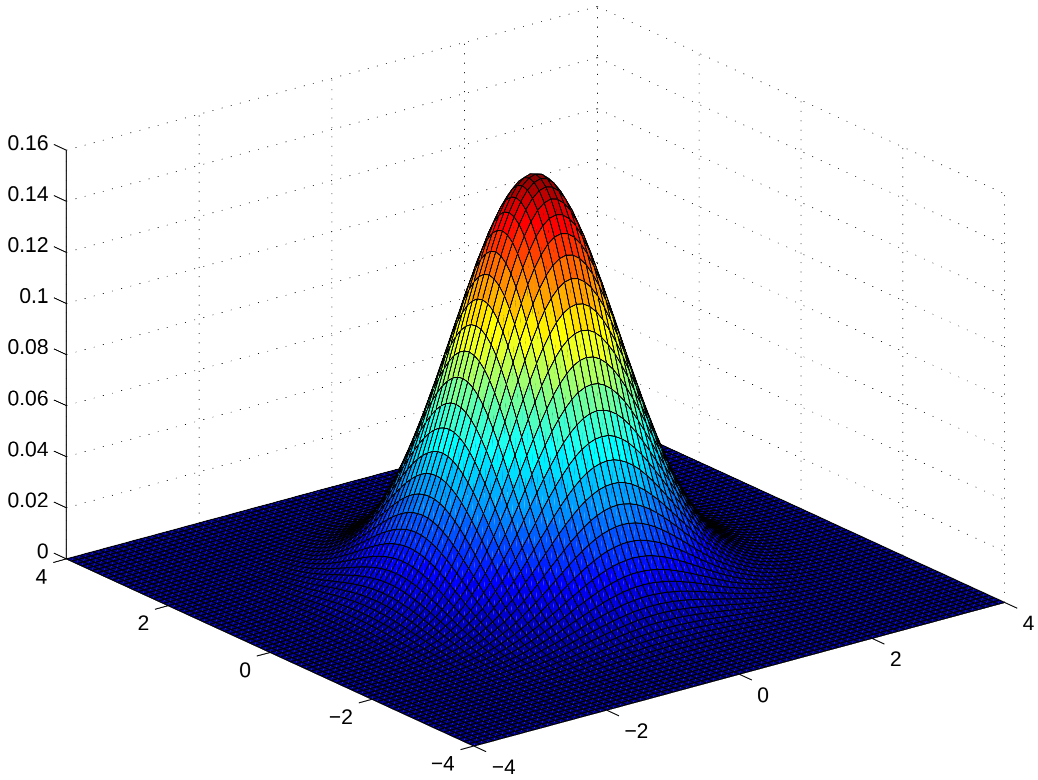
- The multivariate normal distribution function is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

- The mean vector is $\boldsymbol{\mu}$
- The covariance matrix is $\mathbf{\Sigma}$
- Standard notation for multivariate normal distributions is $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$
- Visualizing the MVN is difficult for more than two dimensions, so I will demonstrate some plots with two variables - the bivariate normal distribution

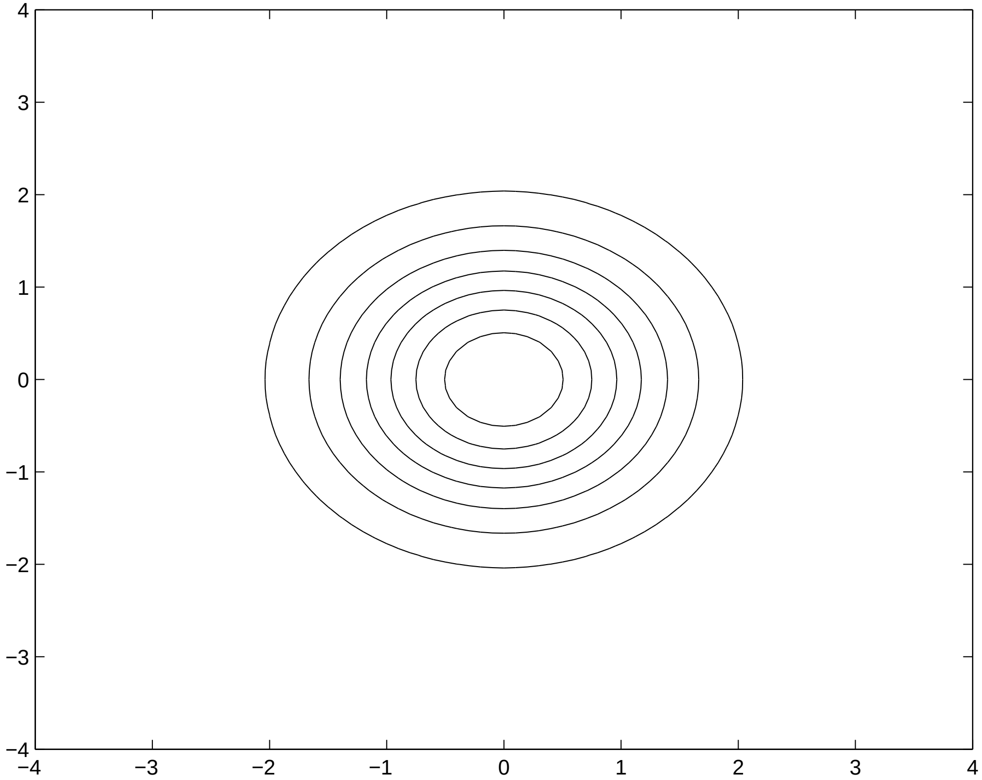
Bivariate Normal Plot #1

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



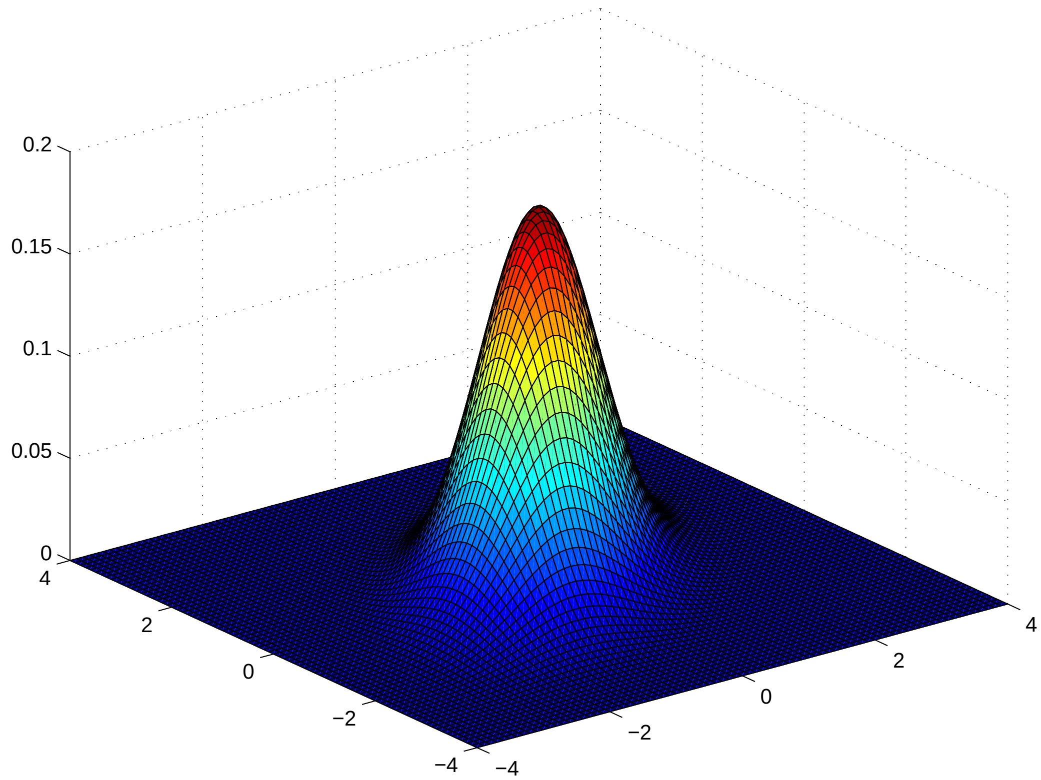
Bivariate Normal Plot #1a

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



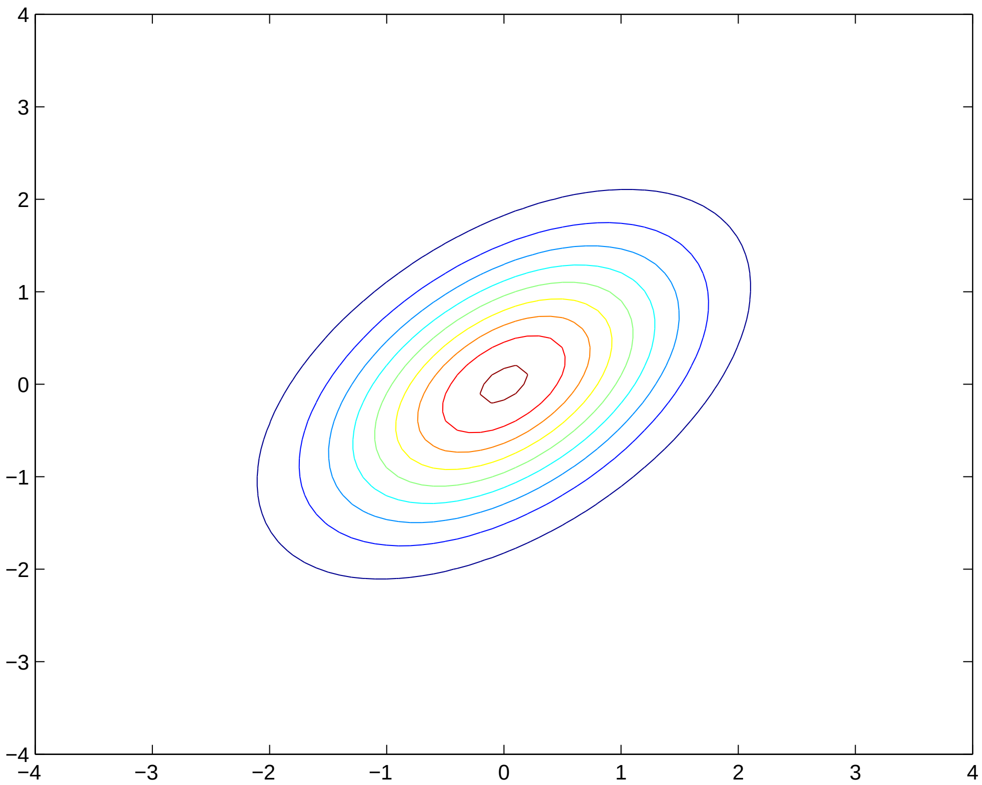
Bivariate Normal Plot #2

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



Bivariate Normal Plot #2

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



MVN Contours

- The lines of the contour plots denote places of equal probability mass for the MVN distribution
  - The lines represent points of both variables that lead to the same height on the z-axis (the height of the surface)
- These contours can be constructed from the eigenvalues and eigenvectors of the covariance matrix
  - The direction of the ellipse axes are in the direction of the eigenvalues
  - The length of the ellipse axes are proportional to the constant times the eigenvector
- Specifically:

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$

has ellipsoids centered at $\boldsymbol{\mu}$, and has axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$

MVN Contours, Continued

- Contours are useful because they provide confidence regions for data points from the MVN distribution
- The multivariate analog of a confidence interval is given by an ellipsoid, where c is from the Chi-Squared distribution with p degrees of freedom
- Specifically:

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \chi^2_p(\alpha)$$

provides the confidence region containing $1 - \alpha$ of the probability mass of the MVN distribution

MVN Contour Example

- Imagine we had a bivariate normal distribution with:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- The covariance matrix has eigenvalues and eigenvectors:

$$\boldsymbol{\lambda} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}, \mathbf{E} = \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix}$$

- We want to find a contour where 95% of the probability will fall, corresponding to $\chi^2_2(0.05) = 5.99$

MVN Contour Example

- This contour will be centered at $\boldsymbol{\mu}$
- Axis 1:

$$\boldsymbol{\mu} \pm \sqrt{5.99 \times 1.5} \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} 2.12 \\ 2.12 \end{bmatrix}, \begin{bmatrix} -2.12 \\ -2.12 \end{bmatrix}$$

- Axis 2:

$$\boldsymbol{\mu} \pm \sqrt{5.99 \times 0.5} \begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} -1.22 \\ 1.22 \end{bmatrix}, \begin{bmatrix} 1.22 \\ -1.22 \end{bmatrix}$$

MVN Properties

- The MVN distribution has some convenient properties
- If **X** has a multivariate normal distribution, then:

1. Linear combinations of **X** are normally distributed
2. All subsets of the components of **X** have a MVN distribution
3. Zero covariance implies that the corresponding components are independently distributed
4. The conditional distributions of the components are MVN

Linear Combinations

- If $\mathbf{X} \sim N_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, then any set of q linear combinations of variables $\mathbf{A}_{(q \times p)}$ are also normally distributed as $\mathbf{AX} \sim N_q\left(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A'}\right)$

- For example, let $p = 3$ and Y be the difference between X_1 and X_2. The combination matrix would be

$$\boldsymbol{A} = \left[\begin{array}{ccc})1 & -1 & 0 \end{array}\right]$$

For $\mathbf{X}$

$$\boldsymbol{\mu} = \left[\begin{array}{c} \mu_1 \\ \mu_2 \\ \mu_3 \end{array}\right], \boldsymbol{\Sigma} = \left[\begin{array}{ccc} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{array}\right]$$

For $\mathbf{Y} = \mathbf{AX}$

$$\boldsymbol{\mu}_Y = \left[\begin{array}{c} \mu_1 - \mu_2 \end{array}\right], \boldsymbol{\Sigma}_Y = \left[\begin{array}{c} \sigma_{11} + \sigma_{22} - 2\sigma_{12} \end{array}\right]$$

MVN Properties

- The MVN distribution is characterized by two parameters:
  - The mean vector $\boldsymbol{\mu}$
  - The covariance matrix $\boldsymbol{\Sigma}$
- The maximum likelihood estimates for these parameters are given by:
  - The mean vector: $\bar{\mathbf{x}}' = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\mathbf{X'1}$
  - The covariance matrix
    $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$

Distribution of $\bar{x}$ and S

Recall back in Univariate statistics you discussed the Central Limit Theorem (CLT)

It stated that, if the set of n observations $x_1, x_2, \ldots, x_n$ were normal or not...

- The distribution of $\bar{x}$ would be normal with mean equal to μ and variance σ^2/n
- We were also told that $(n-1)s^2/\sigma^2$ had a Chi-Square distribution with $n-1$ degrees of freedom
- *Note: We ended up using these pieces of information for hypothesis testing such as t-test and ANOVA.*

Distribution of $\bar{\mathbf{x}}$ and S

We also have a Multivariate Central Limit Theorem (CLT)

It states that, if the set of n observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are multivariate normal or not...

- The distribution of $\bar{\mathbf{x}}$ would be normal with mean equal to $\boldsymbol{\mu}$ and variance/covariance matrix $\boldsymbol{\Sigma}/n$
- We are also told that $(n-1)\mathbf{S}$ will have a Wishart distribution, $W_p(n-1, \boldsymbol{\Sigma})$, with $n-1$ degrees of freedom
  - This is the multivariate analogue to a Chi-Square distribution
- *Note: We will end up using some of this information for multivariate hypothesis testing*

Distribution of $\bar{\mathbf{x}}$ and S

- Therefore, let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be independent observations from a population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
- The following are true:
  - $\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)$ is approximately $N_p(\mathbf{0}, \boldsymbol{\Sigma})$
  - $n\left(\mathbf{X} - \boldsymbol{\mu}\right)' \mathbf{S}^{-1}\left(\mathbf{X} - \boldsymbol{\mu}\right)$ is approximately χ^2_p

Sufficient Statistics

- The sample estimates $\bar{\mathbf{X}}$ and $S)$ are sufficient statistics
- This means that all of the information contained in the data can be summarized by these two statistics alone
- This is only true if the data follow a multivariate normal distribution - if they do not, other terms are needed (i.e., skewness array, kurtosis array, etc...)
- Some statistical methods only use one or both of these matrices in their analysis procedures and not the actual data

Density and Likelihood Functions

- The MVN distribution is often the core statistical distribution for a uni- or multivariate statistical technique

- Maximum likelihood estimates are preferable in statistics due to a set of desirable *asymptotic* properties, including:
  - **Consistency:** the estimator converges in probability to the value being estimated
  - **Asymptotic Normality:** the estimator has a normal distribution with a functionally known variance
  - **Efficiency:** no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE

- The form of the MVN ML function frequently appears in statistics, so we will briefly discuss MLE using normal distributions

An Introduction to Maximum Likelihood

- Maximum likelihood estimation seeks to find parameters of a statistical model (mapping onto the mean vector and/or covariance matrix) such that the statistical likelihood function is maximized

- The method assumes data follow a statistical distribution, in our case the MVN

- More frequently, the log-likelihood function is used instead of the likelihood function
  - The “logged” and “un-logged” version of the function have a maximum at the same point
  - The “logged” version is easier mathematically

Maximum Likelihood for Univariate Normal

- We will start with the univariate normal case and then generalize
- Imagine we have a sample of data, X, which we will assume is normally distributed with an unknown mean **but** a known variance (say the variance is 1)
- We will build the maximum likelihood function for the mean
- Our function rests on two assumptions:

1. All data follow a normal distribution
2. All observations are independent

- Put into statistical terms: X is independent and identically distributed (iid) as $N_1(\mu, 1)$

Building the Likelihood Function

- Each observation, then, follows a normal distribution with the same mean (unknown) and variance (1)

- The distribution function begins with the density – the function that provides the normal curve (with (1) in place of σ^2):

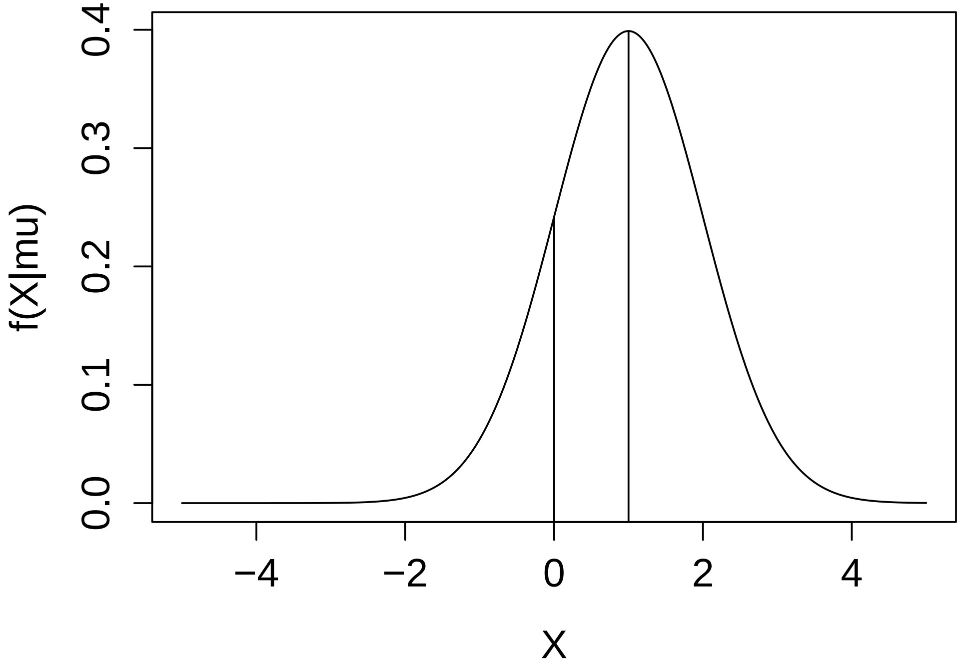
$$f(X_i|\mu) = \frac{1}{\sqrt{2\pi(1)}} \exp\left(-\frac{(X_i-\mu)^2}{2(1)^2}\right)$$

- The density provides the "likelihood" of observing an observation X_i for a given value of μ (and a known value of $\sigma^2 = 1$

- The "likelihood" is the height of the normal curve

The One-Observation Likelihood Function

The graph shows $f(X_i|\mu = 1)$ for a range of X



The vertical lines indicate:

- $f(X_i = 0|\mu = 1) = .241$
- $f(X_i = 1|\mu = 1) = .399$

The Overall Likelihood Function

- Because we have a sample of N observations, our likelihood function is taken across all observations, not just one

- The "joint" likelihood function uses the assumption that observations are independent to be expressed as a product of likelihood functions across all observations:

$$L(\mathbf{x}|\mu) = f(X_1|\mu) \times f(X_2|\mu) \times \ldots \times f(X_N|\mu)$$

$$L(\mathbf{x}|\mu) = \prod_{i=1}^{N} f(X_i|\mu) = \left(\frac{1}{2\pi(1)}\right)^{N/2} \exp\left(-\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{2(1)^2}\right)$$

- The value of μ that maximizes $f(\mathbf{x}|\mu)$ is the MLE (in this case, it's the sample mean)

- In more complicated models, the MLE does not have a closed form and therefore must be found using numeric methods

The Overall Log-Likelihood Function

- For an unknown mean μ and variance σ^2, the likelihood function is:

$$L(\mathbf{x}|\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{2\sigma^2}\right)$$

- More commonly, the log-likelihood function is used:

$$L(\mathbf{x}|\mu,\sigma^2) = -\left(\frac{N}{2}\right)\log\left(2\pi\sigma^2\right) - \left(\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{2\sigma^2}\right)$$

The Multivariate Normal Likelihood Function

- For a set of N independent observations on p variables, $\mathbf{X}_{(N \times p)}$, the multivariate normal likelihood function is formed by using a similar approach
- For an unknown mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the joint likelihood is:

$$L(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\left(\mathbf{x}_i - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)/2\right)$$

$$= \frac{1}{(2\pi)^{np/2}} \frac{1}{|\Sigma|^{n/2}} \exp\left(-\sum_{i=1}^{N} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)/2\right)$$

The Multivariate Normal Likelihood Function

- Occasionally, a more intricate form of the MVN likelihood function shows up
- Although mathematically identical to the function on the last page, this version typically appears without explanation:

$$L(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-np/2} |\Sigma|^{-n/2}$$

$$\exp\left(-\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\right)\right]/2\right)$$

The MVN Log-Likelihood Function

- As with the univariate case, the MVN likelihood function is typically converted into a log-likelihood function for simplicity
- The MVN log-likelihood function is given by:

$$l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\left(\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right)$$

But...Why?

- The MVN distribution, likelihood, and log-likelihood functions show up frequently in statistical methods

- Commonly used methods rely on versions of the distribution, methods such as:
  - Linear models (ANOVA, Regression)
  - Mixed models (i.e., hierarchical linear models, random effects models, multilevel models)
  - Path models/simultaneous equation models
  - Structural equation models (and confirmatory factor models)
  - Many versions of finite mixture models

- Understanding the form of the MVN distribution will help to understand the commonalities between each of these models

MVN in Mixed Models

- From SAS' manual for *proc mixed*:

Estimating Covariance Parameters in the Mixed Model

Estimation is more difficult in the mixed model than in the general linear model. Not only do you have $\boldsymbol{\beta}$ as in the general linear model, but you hav unknown parameters in $\boldsymbol{\gamma}$, $\mathbf{G}$, and $\mathbf{R}$ as well. Least squares is no longer the best method. ***Generalized least squares*** (GLS) is more appropriate, minimizing

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

However, it requires knowledge of $\mathbf{V}$ and, therefore, knowledge of $\mathbf{G}$ and $\mathbf{R}$. Lacking such information, one approach is to use ***estimated*** GLS, in w you insert some reasonable estimate for $\mathbf{V}$ into the minimization problem. The goal thus becomes finding a reasonable estimate of $\mathbf{G}$ and $\mathbf{R}$.

In many situations, the best approach is to use ***likelihood-based*** methods, exploiting the assumption that $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are normally distributed (Hartl Rao 1967; Patterson and Thompson 1971; Harville 1977; Laird and Ware 1982; Jennrich and Schluchter 1986). PROC MIXED implements two like based methods: ***maximum likelihood*** (ML) and ***restricted/residual maximum likelihood*** (REML). A favorable theoretical property of ML and RE that they accommodate data that are missing at random (Rubin 1976; Little 1995).

PROC MIXED constructs an objective function associated with ML or REML and maximizes it over all unknown parameters. Using calculus, it is possible to reduce this maximization problem to one over only the parameters in $\mathbf{G}$ and $\mathbf{R}$. The corresponding log-likelihood functions are as follow

$$\text{ML}: \quad l(\mathbf{G},\mathbf{R}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{n}{2}\log(2\pi)$$

$$\text{REML}: \quad l_R(\mathbf{G},\mathbf{R}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{n-p}{2}\log(2\pi)\}$$

MVN in Structural Equation Models

- From SAS' manual for *proc calis*:

For normal-theory maximum likelihood estimation, the function is

$$F_{ML} = Tr(\mathbf{S}\mathbf{C}^{-1}) - n + ln(det(\mathbf{C})) - ln(det(\mathbf{S}))$$

- This package uses only the covariance matrix, so the form of the likelihood function is phrased using only the Wishart Distribution:

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

Assessing Normality

- Recall from earlier that IF the data have a Multivariate normal distribution then all of the previously discussed properties will hold

- There are a host of methods that have been developed to assess multivariate normality - just look in Johnson & Wichern

- Given the relative robustness of the MVN distribution, I will skip this topic, acknowledging that extreme deviations from normality will result in poorly performing statistics

- More often than not, assessing MV normality is fraught with difficulty due to sample-estimated parameters of the distribution

Transformations to Near Normality

- Historically, people have gone on an expedition to find a transformation to near-normality when learning their data may not be MVN
- Modern statistical methods, however, make that a very bad idea
- More often than not, transformations end up changing the nature of the statistics you are interested in forming
- Furthermore, not all data need to be MVN (think **conditional** distributions)

Final Thoughts

- The multivariate normal distribution is an analog to the univariate normal distribution
- The MVN distribution will play a large role in the upcoming weeks
- We can finally put the background material to rest, and begin learning some statistics methods
- Tomorrow: lab with SAS - the “fun” of *proc iml*
- Up next week: Inferences about Mean Vectors and Multivariate ANOVA