STAT 530/J530
September 6th, 2005

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu

Today

- Missing Data
- The Multivariate Normal Distribution

What if Data is Missing?

- Why is the Data Missing?

*Missing Completely at Random

*Missing at Random

*Unignorable

What if Data is Missing?

- Solutions
  * Deletion
  * Substitution
  * Multiple Imputation
  * Maximum Likelihood

(Multivariate) Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

$$f(\mathrm{x}) = |2\pi\Sigma|^{-1/2} \, e^{-\frac{1}{2}\left((\mathrm{x}-\mu)^T \Sigma^{-1}(\mathrm{x}-\mu)\right)}$$

What About the Covariance?

Just like σ must be greater than 0 for the normal distribution, the covariance Σ must be positive definite for a multivariate normal.

Σ is positive definite if $x^T\Sigma x > 0$ for all vectors x that aren't all zero.

Generating MVN Data

```
library(MASS)
mu<-c(5,0,-1)
sigma<-matrix(c(1,0.5,-0.2,
      0.5,1,0,
      -.2,0,4),
       ncol=3,byrow=T)
x<-mvrnorm(n=1000,mu,sigma)
```


Why?!?

Being positive definite guarantees that:

Σ^{-1} exists and is positive definite

$(x-\mu)^T \Sigma^{-1} (x-\mu)$ is positive

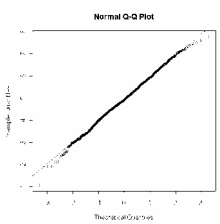
$|\Sigma|$ is positive

$$f(x) = |2\pi\Sigma|^{-1/2} e^{-\frac{1}{2}\left((x-\mu)^T \Sigma^{-1}(x-\mu)\right)}$$

Properties of the MVN

Cross Sections: All linear combinations of normal random variables (including the "marginal distributions" are normally distributed.

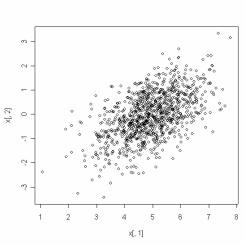
```
qqnorm(x[,1])
qqline(x[,1])
qqnorm(x%*%c(1,2,4))
qqline(x%*%c(1,2,4))
```



Geometry of the Multivariate Normal

"Topographic Maps": The pdf of a multivariate normal makes ellipses of equal probability.

```
plot(x[,1],x[,2])
```



Geometry of the MVN

- Relationships: If the x are random vectors from the multivariate normal distribution with mean vector μ and covariance Σ then: $(x-\mu)^T \Sigma^{-1} (x-\mu) \sim \chi^2_{df=q}$
- This holds approximately if $\overline{X}$ and S are used instead.

Chi-Square Plot

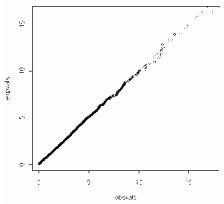
Find the distances…

```
xbar<-apply(x,2,mean)
S<-var(x)
ds<-rep(0,1000)
for (i in 1:1000){
  ds[i]<-t(x[i,]-xbar)%*%
      solve(S)%*%(x[i,]-xbar)}
```

Chi-Square Plot

Make the plot…

```
obsvals<-sort(ds)
expvals<-qchisq((1:1000)/1001,3)
plot(obsvals,expvals)
lines(c(0,20),c(0,20))
```



Homework for Tuesday the 13th

0) Read Chapter 1

1) Imagine that someone wanted to come up with a total score to summarize each persons view of the oil crisis (Q1-Q25).

a) Explain why it doesn't make sense to just add up all of the numbers

b) Find the correlation matrix for Q1-Q25 data set and suggest two separate groups of questions that might be added separately.

Homework for Tuesday the 13th

c) How could these two scores be combined to form a single score?

2) Check whether the data set **`normsamp.txt`** on the web is actually multivariate normal. (Code will be posted shortly for the various checks we went through today).