Statistics Basics

Liming Feng

Dept. of Industrial & Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Basic techniques

- Descriptive statistics
- Parameter estimation
- Hypothesis testing
- Resampling
- Modeling univariate and multivariate data

Basic descriptive tools

- Organize and summarize data, describe the main features
- Compute **summary statistics**: sample mean, variance, skewness, kurtosis, covariance/correlation, etc.
- **Visualize data**: histogram, time series plot, box plot, scatter plot, probability plot

1. Computing summary statistics

Ruppert 2011 Chapter 4

Sample mean/variance/stdev

- Sample **mean** of data set $\{x_1, \cdots, x_n\}$: measures location of the data

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Sample **variance/standard deviation**: measure variability of the data

$$\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

- We use x_i to denote a numeric value in the above, and X_i to denote a r.v.

Sample covariance/correlation/skewness/kurtosis

- Sample **covariance/correlation** of data set $\{(x_i, y_i) : 1 \le i \le n\}$: measure linear relationship

$$\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n), \quad \hat{\rho}_{xy} = \hat{\sigma}_{xy} / (\hat{\sigma}_x \hat{\sigma}_y)$$

where $\hat{\sigma}_x, \hat{\sigma}_y$ are sample standard deviations

- Sample **skewness/kurtosis**: measure skewness and tail fatness

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right)^{3/2}}, \quad \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right)^{2}}$$

Entering data in R

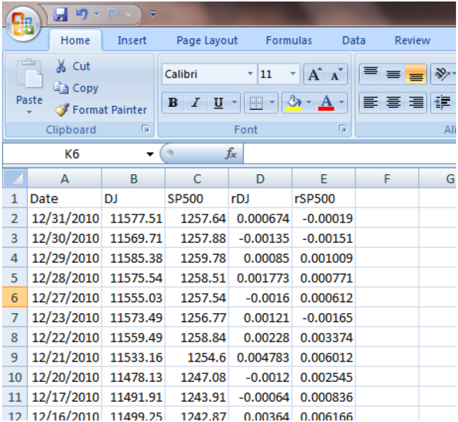
- Small data size

  stockreturn <- c(0.01, -0.02, 0.01, 0.01, 0.00)

  stock return is a vector
- Large data size
  - Save data as a .txt file in your working directory (set this from R file menu or use *setwd*) and read using *read.table*
  - Or save data as a .csv file and read using *read.csv*
  - Get help using *help (read.csv)*

DJIA and SP500

- Daily (adjusted) prices from 1/2/2009 to 12/31/2010; daily log return = log(price today/price previous day); saved as dj-sp500.csv



	A	B	C	D	E	F	G
1	Date	DJ	SP500	rDJ	rSP500		
2	12/31/2010	11577.51	1257.64	0.000674	-0.00019		
3	12/30/2010	11569.71	1257.88	-0.00135	-0.00151		
4	12/29/2010	11585.38	1259.78	0.00085	0.001009		
5	12/28/2010	11575.54	1258.51	0.001773	0.000771		
6	12/27/2010	11555.03	1257.54	-0.0016	0.000612		
7	12/23/2010	11573.49	1256.77	0.00121	-0.00165		
8	12/22/2010	11559.49	1258.84	0.00228	0.003374		
9	12/21/2010	11533.16	1254.6	0.004783	0.006012		
10	12/20/2010	11478.13	1247.08	-0.0012	0.002545		
11	12/17/2010	11491.91	1243.91	-0.00064	0.000836		
12	12/16/2010	11499.25	1242.87	0.00364	0.006166		

- *read.csv* creates a data frame: index

```
> setwd("c:/Liming Feng/Teaching/13-IE522-F12/R")
> index <- read.csv("dj-sp500.csv",header=T)
> index
          Date       DJ   SP500       rDJ    rSP500
1   12/31/2010 11577.51 1257.64  0.000674 -0.000191
2   12/30/2010 11569.71 1257.88 -0.001353 -0.001509
3   12/29/2010 11585.38 1259.78  0.000850  0.001009
4   12/28/2010 11575.54 1258.51  0.001773  0.000771
5   12/27/2010 11555.03 1257.54 -0.001596  0.000612
6   12/23/2010 11573.49 1256.77  0.001210 -0.001646

                            ... ...

499   1/9/2009  8599.18  890.35 -0.016525 -0.021533
500   1/8/2009  8742.46  909.73 -0.003111  0.003391
501   1/7/2009  8769.70  906.65 -0.027598 -0.030469
502   1/6/2009  9015.10  934.70  0.006925  0.007787
503   1/5/2009  8952.89  927.45 -0.009095 -0.004679
504   1/2/2009  9034.69  931.80        NA        NA
>
```

Compute sample mean/stdev/correlation in R

- Use \$ sign to access a variable in a data frame, use "na.rm=T" to remove NA's before computing a sample mean

```
> mean(index$DJ)
[1] 9777.127
> mean(index$rDJ)
[1] NA
> mean(index$rDJ,na.rm=T)
[1] 0.0004930199
> sapply(index,mean,na.rm=T)
         Date           DJ        SP500          rDJ       rSP500
           NA 9.777127e+03 1.044006e+03 4.930199e-04 5.961869e-04
```

- Sample standard deviations: *sd(index\$rDJ,na.rm=T)* returns 1.29%. This is the sample standard deviation of daily log returns in DJIA
- *cor(index\$rDJ,index\$rSP500,use="complete.obs")* returns a sample correlation of 98.48%, indicating a strong positive linear relationship

Install R libraries

- Install *moments* library: Packages → Install package(s) → select CRAN → select package *moments*
- Call the library so that you can use the functions in the library

```
> library(moments)
> skewness(index$rSP500,na.rm=T)
[1] -0.09232887
> kurtosis(index$rSP500,na.rm=T)
[1] 5.655863
```

- Sample skewness of rSP500: -0.092 (negatively skewed)
- Sample kurtosis of rSP500: 5.656 (fatter tails than normal)

Quantiles

- **Sample median** (0.5-quantile) divides data into two halves
- **First quartile** q_1 (0.25-quantile): approximately 25% of observations are below the first quartile, 75% are above
- **Third quartile** q_3 (0.75-quantile): 75% are below the third quartile, 25% are above
- Interquartile range (**IQR**) $q_3 - q_1$: another commonly used measure of variability

```
> quantile(index$rSP500,na.rm=T)
        0%        25%        50%        75%       100%
-0.0542620 -0.0054410  0.0011240  0.0073025  0.0683660
> IQR(index$rSP500,na.rm=T)
[1] 0.0127435
```

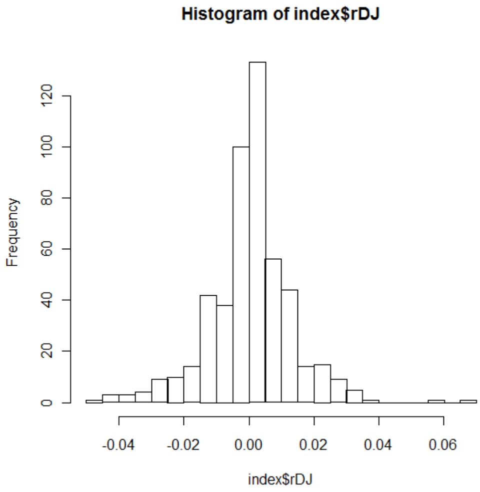
2. Visualize Data

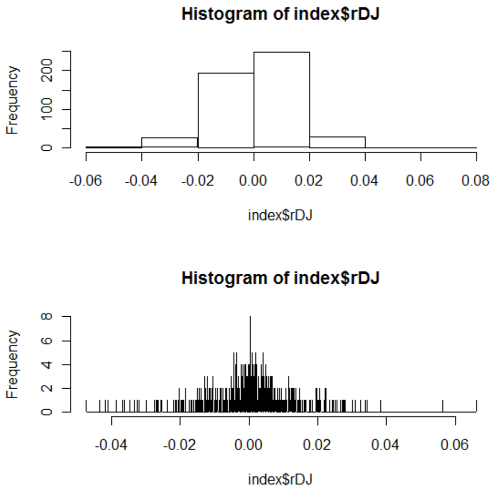
Ruppert 2011 Chapter 4

Histogram

- Visualize the shape of the distribution
- Divide the range of the data into small bins, count the number of observations in each bin (**frequency**)
- Number of bins specified by *breaks=k* (k bins) in R
- k should not be too large or too small

```
> hist(index$rDJ,breaks=30)
> par(mfrow = c(2,1))
> hist(index$rDJ,breaks=5)
> hist(index$rDJ,breaks=1000)
>
```


Histogram of index$rDJ
Frequency
0
20
40
60
80
100
120
-0.04
-0.02
0.00
0.02
0.04
0.06
index$rDJ



Histogram of index$rDJ
Frequency
0
100
200
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
index$rDJ
Histogram of index$rDJ
Frequency
0
2
4
6
8
-0.04
-0.02
0.00
0.02
0.04
0.06
index$rDJ

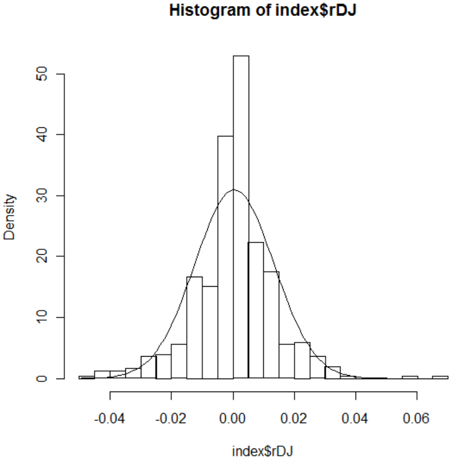
Histogram with a normal fit

- Make the vertical axis of the histogram **density** instead of **frequency**
- Add a fitted normal pdf on the histogram

```
> hist(index$rDJ,breaks=30,prob=T)
> rDJmean=mean(index$rDJ,na.rm=T);
> rDJsd=sd(index$rDJ,na.rm=T);
> curve(dnorm(x,rDJmean,rDJsd),add=T,col="grey",lwd=4)
```

  lwd=4 specifies line thickness
- *dnorm* computes the pdf of a normal distribution
- With *add=T*, the new curve will be added to an existing plot

- Data exhibit fatter tails than normal distribution (normal distribution understates likelihood of extreme price movements; kurtosis can be computed to be 5.9)

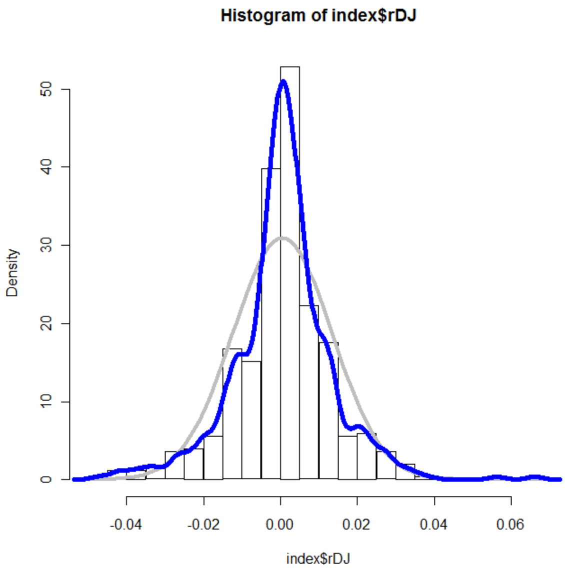


Kernel density estimation

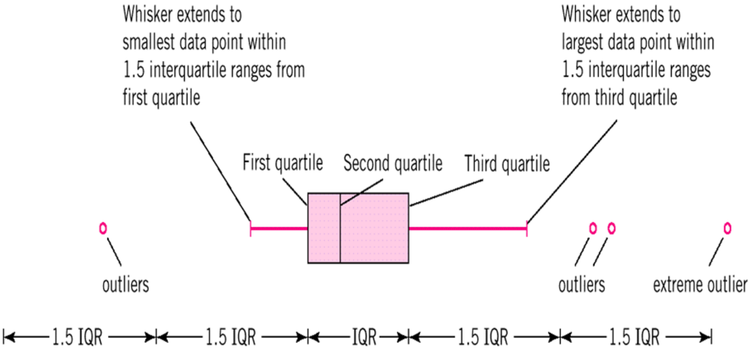
- Given a data set $\{x_1, \cdots, x_n\}$, the **kernel density estimate** (KDE) is given by

$$\hat{f}(x) = \frac{1}{nb}\sum_{i=1}^{n} K((x - x_i)/b)$$

  The **kernel function** K often takes standard normal density
- Then $\frac{1}{b}K((x - x_i)/b)$ is the density of $N(x_i, b^2)$; it is thin around x_i when the **bandwidth** b is small. b determines the resolution of KDE; R determines b automatically
- *lines(density(index$rDJ,na.rm=T),col="blue",lwd=5)* (*lines* plots the KDE and adds to the existing plot)

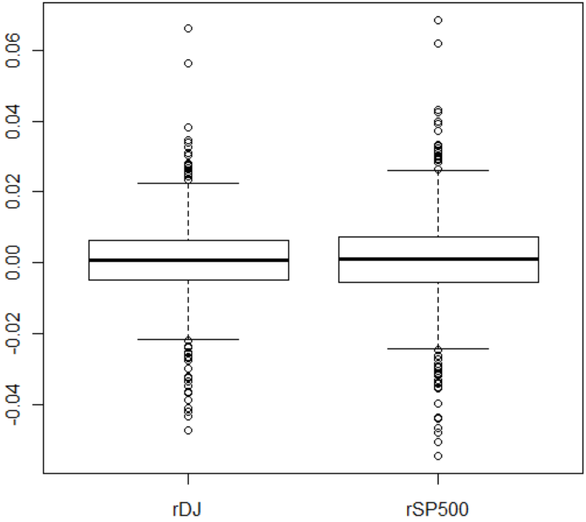

Histogram of index$rDJ
Density
0
10
20
30
40
50
-0.04
-0.02
0.00
0.02
0.04
0.06
index$rDJ

Box plot



Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
outliers
outliers
extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

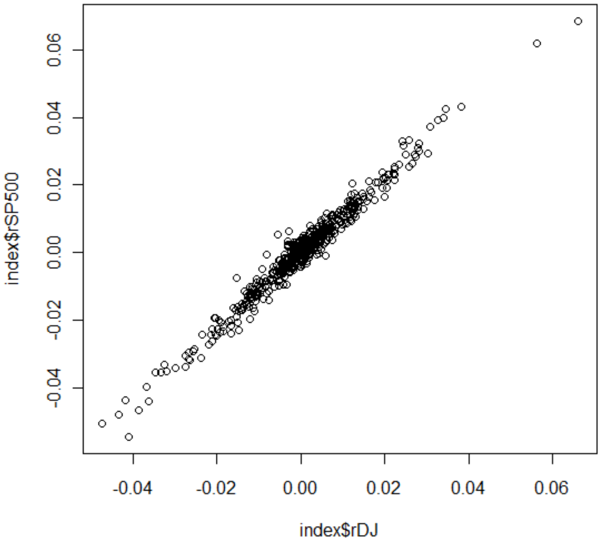
- Box plot visualizes the location (median) and variability (IQR) of the data
- Identifies possible outliers (extreme observations)
- Outliers can be legitimate data and shouldn't be automatically discarded
- Useful for side-by-side comparison of data sets
- *boxplot(index$rDJ,index$rSP500,names=c("rDJ","rSP500"))*


0.06
0.04
0.02
0.00
-0.02
-0.04
rDJ
rSP500

Bivariate data

- Linear relationship between returns of DJIA and SP500
- Recall that the sample correlation coefficient is 98.47669%: strong positive linear relationship
- Visualize the linear relationship using **scatter plot**:
  plot(index$rDJ,index$rSP500)

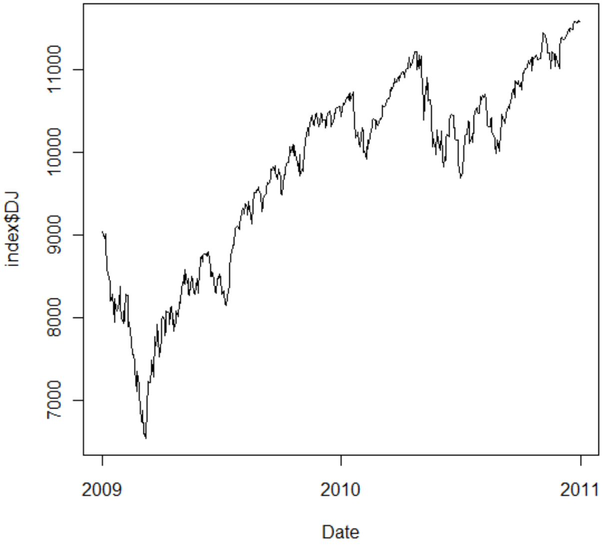
Scatter plot



Time series plot

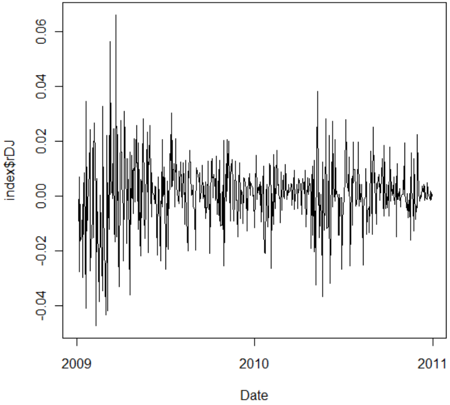
- Financial data are time series data
- Time series plot visualizes **dynamics** of a financial variable
- Identify **cycles, trends, anomalies**
- Convert the first column of the data frame *index* to a date object:
  Date <- as.Date(index$Date,"%m/%d/%Y")
- *plot(Date,index$DJ,type="l")* (here "l" refers to line)

Price process



Return process

- Repeat the above for the daily return of DJIA



- **Volatility clustering**: large price movements followed by more large price movements

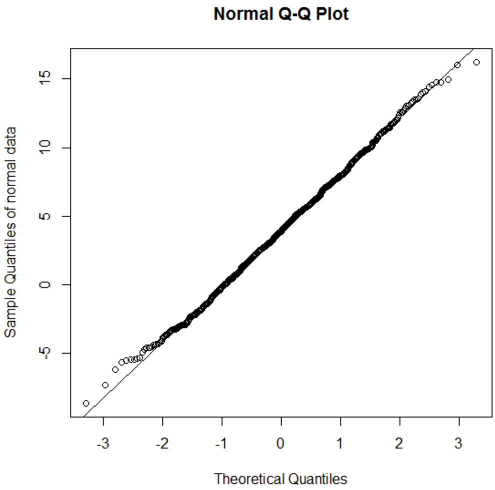
Probability plot (Q-Q plot)

- Visualize whether a certain distribution adequately model a financial variable
- Compare quantiles of the proposed distribution and data
- $q-$quantile of a strictly increasing distribution F is $F^{-1}(q)$: probability of observing a value below $F^{-1}(q)$ is q
- $q-$quantile of a sample $\{x_1, \cdots, x_n\}$: divide data into the lower $100q\%$ and upper $100(1-q)\%$
- **Q-Q plot**: plot theoretical quantiles against sample quantiles; expect a straight line if the data are from the proposed distribution

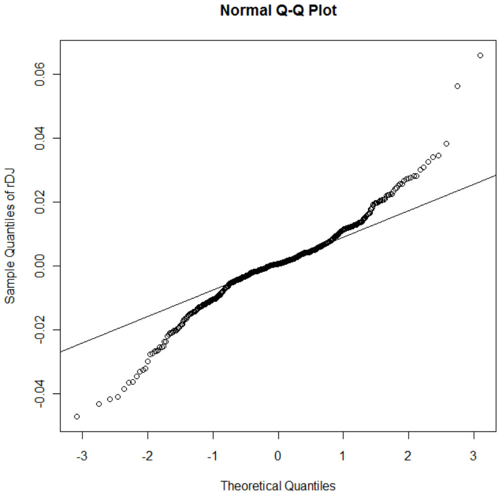
Normal probability plot

- Check whether data follow $N(\mu, \sigma^2)$: sample quantiles $\approx \mu + \sigma\Phi^{-1}(q)$?
- Do sample quantitles and $\Phi^{-1}(q)$ show a linear relationship?
- Shape of the plot indicates type of deviation from normality
- It is essential that you understand which axis contains the data
- In R, y-axis contains the data by default (if you want data on the x-axis as in the book, use *datax=T*):

```
> normaldata <- rnorm(1000,mean=4,sd=4)
> qqnorm(normaldata,ylab = "Sample Quantiles of normal data")
> qqline(normaldata)
> qqnorm(index$rDJ,ylab = "Sample Quantiles of rDJ")
> qqline(index$rDJ)
```


Normal Q-Q Plot
Sample Quantiles of normal data
Theoretical Quantiles
-5
0
5
10
15
-3
-2
-1
0
1
2
3

- Return in DJIA exhibit more extreme observations (on both tails) and suggest distributions with fatter tails



Q-Q plot with Laplace distribution

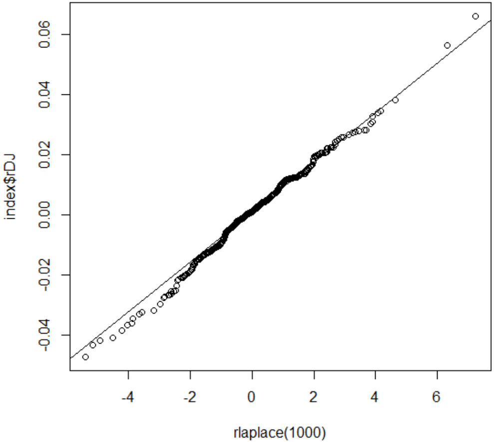
- Recall a Laplace distribution with density

$$p(x) = \frac{1}{2b} e^{-|x-\mu|/b},$$

  mean μ, standard deviation $\sqrt{2}b$, kurtosis 6. It has fatter tails

- *library(VGAM)*
  qqplot(rlaplace(1000),index$rDJ)
  qqline(index$rDJ)
  Here x-axis contains sample quantiles of 1000 simulated observations from the Laplace distribution

- The Laplace distribution captures the tails better



Scatterplot matrix for multivariate data

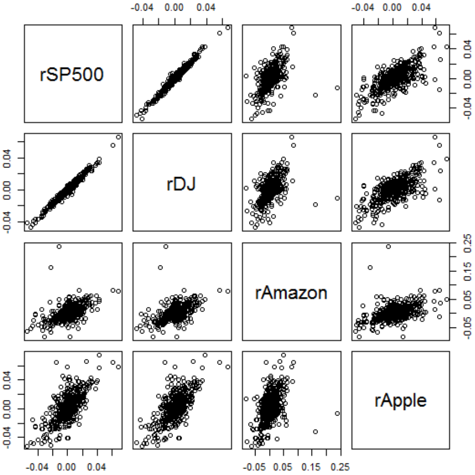
- Covariance and correlation measure linear relations
- Scatterplot matrix visualizes linear (and possibly nonlinear!) relations among several variables: a matrix of plots with ijth entry the scatterplot of the ith and the jth variables
- Daily log returns of S&P500, DJIA, Amazon, Apple (prices from 1/2/2009 to 12/31/2010)

	Date	rSP500	rDJ	rAmazon	rApple
1	12/31/2010	-0.000191	0.000674	-0.015162248	-0.003404417
2	12/30/2010	-0.001509	-0.001353	-0.003386871	-0.005023510
……					
501	1/7/2009	-0.030469	-0.027598	-0.020430439	-0.021845133
502	1/6/2009	0.007787	0.006925	0.059252655	-0.016631513
503	1/5/2009	-0.004679	-0.009095	-0.005534048	0.041337564

```
data=read.csv("stocks.csv",header=T)
stock=as.matrix(data[,2:5])
pairs(stock)
cor(stock)
```

Scatterplot matrix

- Asymmetric: 3rd row 1st column: x axis rSP500, y axis rAmazon; 1st row 3rd column: x axis rAmazon, y axis rSP500



- Sample correlation matrix

	rSP500	rDJ	rAmazon	rApple
rSP500	1.0000000	0.9847669	0.5053974	0.7030486
rDJ	0.9847669	1.0000000	0.4880029	0.6807942
rAmazon	0.5053974	0.4880029	1.0000000	0.4460371
rApple	0.7030486	0.6807942	0.4460371	1.0000000

3. Point estimation

Reference: Ruppert 2011, Appendix A

Random sample

- Interested in parameters of a **population**: draw a random sample and construct an estimator
- A **random sample** of size n from a probability distribution (the population) are n independent r.v.'s $\{X_1, \cdots, X_n\}$, each of them having that distribution (i.i.d.)
- Denote the numerical realization of a random sample by $(x_1, \cdots, x_n)$
- **Statistic**: a function of $\{X_1, \cdots, X_n\}$

Point estimation

- **Point estimator**: estimate a parameter θ of a population by a statistic $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ (**point estimate**: numerical realization $\hat{\theta}(x_1, \cdots, x_n)$)
- Estimate the population mean μ with the **sample mean**

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

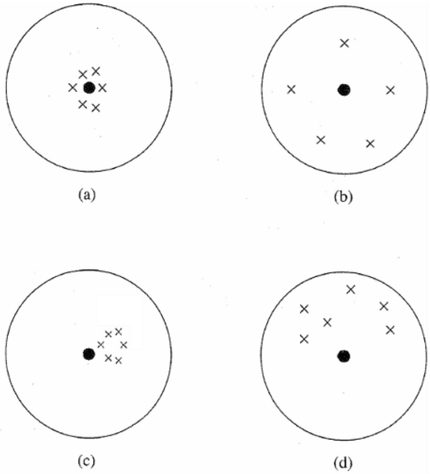
- Estimate the population variance σ^2 with the **sample variance**

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Accuracy and precision

Accurate
Precise

Accurate
Not precise



Precise
Not accurate

Neither

Sampling distribution

- Measure accuracy by $\mathbb{E}[\hat{\theta}] - \theta$
- Measure precision by the variance of $\hat{\theta}$
- Probability distribution of $\hat{\theta}$ is called **sampling distribution**
- Consider the sample mean $\bar{X}_n$

$$\mathbb{E}[\bar{X}_n] = \mu, \; \mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

  where μ, σ^2 are the population mean and variance
- Consider a normal population $N(\mu, \sigma^2)$ and a random sample $\{X_1, \cdots, X_n\}$. Sampling distribution of $\bar{X}_n$

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

Law of large numbers (LLN)

- For larger sample size n, the variance of $\bar{X}_n$ is closer to 0. $\bar{X}_n$ gets "closer" to μ
- **Law of large numbers**: for a random sample $\{X_1, X_2, \cdots\}$ from a distribution with finite mean μ,

$$\lim_{n \to +\infty} \bar{X}_n = \mu \quad a.s.$$

- For a normal population, the convergence is of the order $1/\sqrt{n}$

$$\bar{X}_n - \mu \sim \frac{\sigma}{\sqrt{n}} N(0,1)$$

- What's the "convergence rate" for an arbitrary population?

Central limit theorem (CLT)

- **Central limit theorem**: for a random sample $\{X_1, X_2, \cdots\}$ from a distribution with finite mean μ and variance σ^2,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \Rightarrow N(0,1), \quad n \to \infty$$

- For arbitrary population (even discrete), $\bar{X}_n \sim N(\mu, \sigma^2/n)$ approximately
- Let $\{X_1, X_2, \cdots\}$ be i.i.d. positive r.v.'s. What's the asymptotic distribution of the geometric average $\tilde{X}_n := (X_1 X_2 \cdots X_n)^{1/n}$?

Bias and accuracy

- The **bias** of an estimator is given by $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$
- $\hat{\theta}$ is **unbiased** estimator of θ if $\mathbb{E}[\hat{\theta}] = \theta$
- Sample mean is unbiased: $\mathbb{E}[\bar{X}_n] = \mu$
- It can be shown that **sample variance is unbiased**

  $$\mathbb{E}[S_n^2] = \sigma^2$$

  On the contrary, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is slightly biased
- Sample standard deviation S_n is usually biased

Standard error and precision

- **Standard error** $se(\hat{\theta})$ = standard deviation of $\hat{\theta}$
- Compare $\bar{X}_n$ and $\tilde{X} = (X_1 + X_n)/2$ when $n > 2$: both are unbiased estimators of μ. But

  $$se(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}, se(\tilde{X}) = \frac{\sigma}{\sqrt{2}}$$

  $\tilde{X}$ is less precise with larger standard error
- **MVUE** desired when possible: **m**inimum **v**ariance **u**nbiased **e**stimator; difficult except in special cases (e.g., for normal population, sample mean is MVUE)

MSE

- Balance between accuracy and precision
- **Mean square error** (MSE)

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = (se(\hat{\theta}))^2 + (\text{bias}(\hat{\theta}))^2$$

- Between two estimators: if both unbiased, select the one with smaller standard error; if at least one is biased, compare MSE

An illustration

- Consider a random sample of size n from a Laplace distribution with $\mu = 2, b = \sqrt{2}$ (mean $\mu = 2$, standard deviation $\sigma = 2$)
- Illustrating the law of large numbers

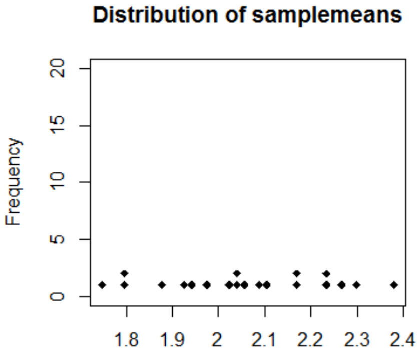
```
> library(VGAM)
> LaplaceData <- rlaplace(10000,location=2,scale=sqrt(2))
> mean(LaplaceData[1:100])
[1] 1.780082
> mean(LaplaceData[1:1000])
[1] 1.920509
> mean(LaplaceData[1:10000])
[1] 1.994072
```

- As n increases from 100 to 1000 to 10,000, $\bar{x}_n$ gets closer to $\mu = 2$

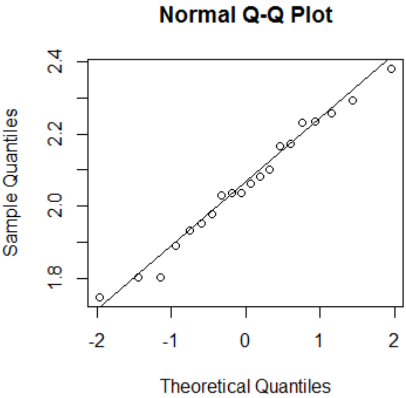
- Illustrating the central limit theorem: draw a sample of size $n = 100$, compute sample mean, repeat for $m = 20$ times; standard error $= \sigma/\sqrt{n} = 0.2$

```
> n=100
> m=20
> LaplaceData <- rlaplace(n*m,location=2,scale=sqrt(2))
> samplemeans <- vector("numeric", length=m)
> for (i in 1:m){samplemeans[i] <- mean(LaplaceData[((i-1)*n+1):(i*n)])}
> library(epicalc)
> dotplot(samplemeans)
> qqnorm(samplemeans)
> qqline(samplemeans)
> sd (samplemeans) # this is the estimated standard error, should be around 0.2
[1] 0.1863437
```

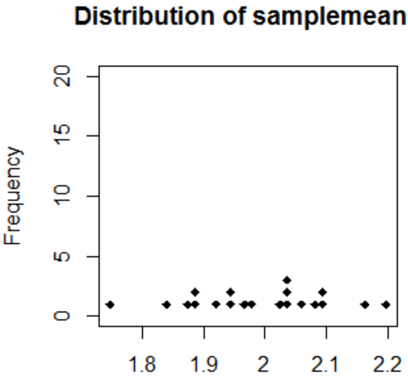
- Dotplot when $n = 100$ (each dot represents one sample mean)



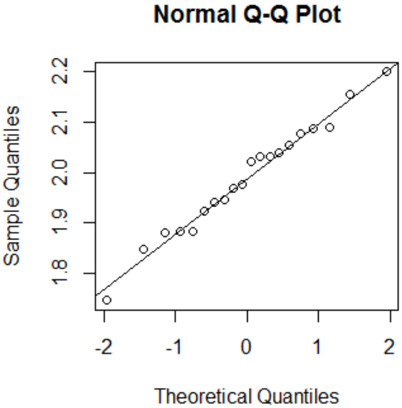
- Normal plot ($n = 100$) illustrates that $\bar{X}_n$ is roughly normal



- Let $n = 400$, standard error $= \sigma/\sqrt{n} = 0.1$
- Dotplot when $n = 400$ (estimates are now more focused around 2, estimated standard error 0.11)



- Normal plot ($n = 400$)



4. Monte Carlo simulation

Reference: Ross 2010, Sections 11.1, 11.2.1, 11.3.1

Monte Carlo methods

- In financial applications, often need to compute

$$\mu = \mathbb{E}[X]$$

  where $\mathbb{E}[X]$ has no analytical expression
- Simulate i.i.d. copies of X: $\{X_i, i \geq 1\}$. By LLN and CLT,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu, \quad \bar{X}_n - \mu \approx \frac{\sigma}{\sqrt{n}} N(0,1)$$

- Rate of convergence of monte carlo simulation: $1/\sqrt{n}$
- Attractive for high dimensions (where other numerical methods often fail)

- **Simulate a continuous r.v.**

  Let $U \sim U[0,1]$. For a continuous r.v. X with strictly increasing cdf $F(x)$,

$$\begin{aligned}\mathbb{P}(X \leq x) &= F(x) \\ &= \mathbb{P}(U \leq F(x)) \\ &= \mathbb{P}(F^{-1}(U) \leq x)\end{aligned}$$

  So X can be simulated from $F^{-1}(U)$

Simulate exponential r.v.'s

- To simulate an exponential r.v. $X \sim Exp(\lambda)$,

$$F(x) = 1 - e^{-\lambda x} \Rightarrow F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x)$$

- For $U \sim U[0,1]$, X can be generated from $-\frac{1}{\lambda}\ln(1-U)$
- Since $1-U$ is still uniform on $[0,1]$, we can generate X simply from $-\frac{1}{\lambda}\ln(U)$

Simulate normal r.v.'s

- Simulate a standard normal r.v.
  - **Inverse transform method**: $\Phi^{-1}(U)$ where $U \sim U[0,1]$, $\Phi(x)$ is the cdf of $N(0,1)$
  - **Box-Muller**: Suppose $Z_1, Z_2 \sim N(0,1)$ are independent. Let

    $$Z_1 = r\cos(\theta), \quad Z_2 = r\sin(\theta)$$

    Then $r^2 \sim Exp(1/2)$ and $\theta \sim U[0, 2\pi]$, r and θ are independent.

    Box Muller algorithm:

    1. *Simulate two independent uniform r.v.'s:* $U_1, U_2 \sim U[0,1]$
    2. $r = \sqrt{-2\ln(U_1)}, \quad \theta = 2\pi U_2$
    3. $Z_1 = r\cos(\theta), \quad Z_2 = r\sin(\theta)$
- Two copies of $U[0,1]$ produce two copies of $N(0,1)$; repeat

Algorithm

- Since $\sigma^2 = \mathrm{var}(X)$ is unknown but can be estimated from the sample variance s_n^2, we report the **estimated standard error** of the sample mean $s_n/\sqrt{n}$. It can be shown that

$$s_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2\right)$$

- Algorithm to estimate $\mu = \mathbb{E}[X]$ using Monte Carlo simulation
  Generate x_1, let $\bar{x} = x_1, \bar{y} = x_1^2$
  For $k = 2 : n$
      Generate x_k
      Update sample mean: $\bar{x} = (1 - \frac{1}{k})\bar{x} + \frac{1}{k}x_k$
      Update $\bar{y}$: $\bar{y} = (1 - \frac{1}{k})\bar{y} + \frac{1}{k}x_k^2$
  Compute $se = (\frac{1}{n-1}(\bar{y} - \bar{x}^2))^{1/2}$. Report $\bar{x}$ and se

Understanding standard error

- Let z_α be the $1-\alpha$ quantile of $N(0,1)$. By CLT,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

$$\mathbb{P}(|\bar{X}_n - \mu| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \approx 1-\alpha$$

$$\mathbb{P}\left(\mu \in [\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]\right) \approx 1-\alpha$$

- $\alpha = 0.05, z_{\alpha/2} = 1.96$: with probability 95%, the true mean μ is in the interval $\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Illustrating Monte Carlo estimation

- Compute $\mu = \mathbb{E}[\sin(X)]$ where $X \sim N(2, 1)$

```
> n=1000
> x=sin(rnorm(n,mean=2,sd=1))
> muhat=mean(x); muhat
[1] 0.5413254
> se=sd(x)/sqrt(n); se
[1] 0.01532193
> z= -qnorm(0.025)
> muhat-se*z
[1] 0.511295
> muhat+se*z
[1] 0.5713559
```

- The interval $[0.51, 0.57]$ is called the 95% confidence interval for μ
- Is the following statement true: the probability that $\mu \in [0.51, 0.57]$ is 95%

- Each time the above is repeated, one gets a different estimate and CI. 95% of the time, these CI's contain the true mean
- To have a tighter CI, one may increase the sample size to reduce the standard error
- With $n = 1000,000$, I obtain estimate 0.552, se 0.0005, 95% CI $[0.551, 0.553]$
- To improve a standard Monte Carlo estimation, computational cost increases rapidly
- Goals of efficient monte carlo simulation: reducing S_n (**variance reduction techniques**) or increasing order of convergence from $1/\sqrt{n}$ to $1/n$ (**quasi-Monte Carlo**)

5. Gaussian samples

Ruppert 2011, Appendix A.10, A.17

CI for mean: large sample

- For a **large sample** from some population

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

at least approximately. Sample standard deviation S_n can be used in the place of σ

- Approximate $1-\alpha$ confidence interval for the mean

$$\bar{x}_n \pm z_{\alpha/2}\frac{s_n}{\sqrt{n}}$$

- One-sided CI: e.g., approximate $1-\alpha$ **upper confidence bound**

$$\bar{x}_n + z_{\alpha}\frac{s_n}{\sqrt{n}}$$

Chi-squared distribution

- What if the sample size is small? What's the distribution of

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

- Assume a **normal population**
- If $X_1, \cdots, X_\nu \sim N(0,1)$ are independent, then

$$X_1^2 + \cdots + X_\nu^2$$

has a gamma distribution with shape parameters $\nu/2$ and rate parameter $1/2$. We call this distribution a **chi-squared distribution** with ν df (degrees of freedom), denoted by χ^2_ν

- Given a random sample $\{X_1, \cdots, X_n\}$ from $N(\mu, \sigma^2)$
- S_n^2 and $\bar{X}_n$ are independent:

$$\mathrm{cov}(\bar{X}_n, X_i - \bar{X}_n) = \mathrm{cov}(\bar{X}_n, X_i) - \mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$$

  Independence follows since for bivariate normal r.v.'s, zero correlation implies independence
- It can been shown that

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2 + \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)^2$$

$$\chi_n^2 \qquad\qquad \chi_{n-1}^2 \qquad\qquad \chi_1^2$$

- For $X_1, \cdots, X_n$ i.i.d. from a normal distribution

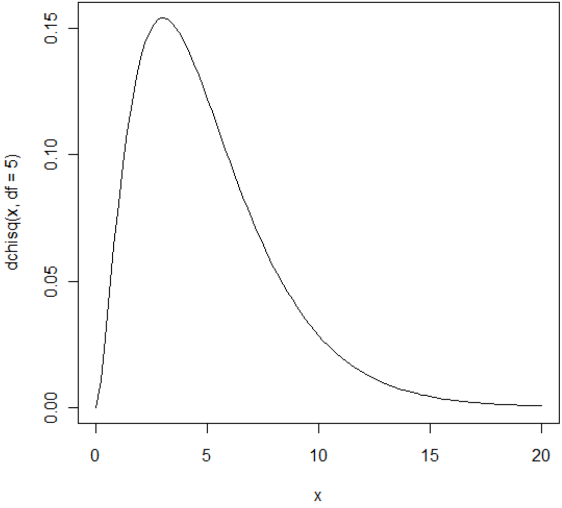
$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

- $1-\alpha$ conference interval for σ

$$\frac{(n-1)S_n^2}{\sigma^2} \in [\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}] \Rightarrow \sigma^2 \in \left[\frac{(n-1)S_n^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$

- $\chi^2_{\alpha/2,n-1}$ is the $1-\alpha/2$ quantile of χ^2_{n-1}; obtained in R as *qchisq(1-α/2,df=n-1)*
- Plot pdf of χ^2_5 in R: *curve(dchisq(x,df=5),from=0,to=20)*

Chi-squared pdf



Student's t

- Suppose $Z \sim N(0,1)$ and $X \sim \chi^2_\nu$ are independent. Then

$$T = \frac{Z}{\sqrt{X/\nu}}$$

has a t **distribution** with ν degrees of freedom, denoted by t_ν. The pdf of t_ν

$$p(x) = \frac{\Gamma((\nu+1)/2)}{(\pi\nu)^{1/2}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \; x \in \mathbb{R}$$

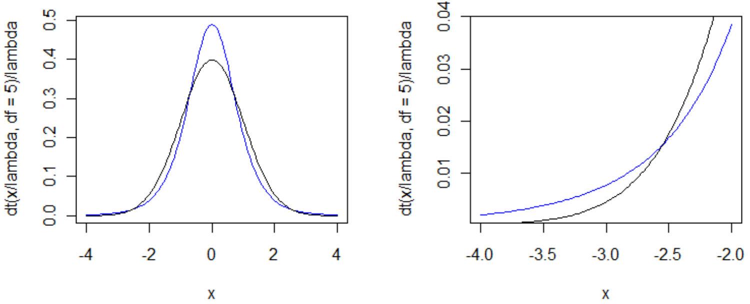
- For a random sample $\{X_1, \cdots, X_n\}$ from a normal population

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}} \sim t_{n-1}$$

t vs normal

- t distributions have heavier tails than normal distributions. Mean exists and equals 0 when $\nu > 1$. Variance is finite and equals $\nu/(\nu-2)$ when $\nu > 2$. Kurtosis is finite and equals $3 + 6/(\nu-4)$ when $\nu > 4$
- As $\nu \to +\infty$, t_ν converges to $N(0,1)$ (by LLN, for $X \sim \chi^2_\nu$, $X/\nu \to 1$)
- If $X \sim t_\nu$, $\mu + \lambda X$ has mean μ and variance $\lambda^2\nu/(\nu-2)$ (for $\nu > 2$)
- Standardization: for $\nu > 2$, let $\mu = 0, \lambda = \sqrt{(\nu-2)/\nu}$, $\mu + \lambda X$ has mean 0 and variance 1

- *lambda=sqrt((5-2)/5)*
  curve(dt(x/lambda,df=5)/lambda,from=-4,to=4,col="blue")
  curve(dnorm(x),from=-4,to=4,add=T)



$t-$intervals

- Let $t_{\alpha,n-1}$ be the $1-\alpha$ quantile. In R, $t_{\alpha,n-1}$ can be computed from $qt(1-\alpha, n-1)$

$$t_{0.025,4} = 2.776,\ \ t_{0.025,100} = 1.984,\ \ z_{0.025} = 1.96$$

- For a random sample $\{X_1, \cdots, X_n\}$ from a normal population

$$\mathbb{P}(|T| \leq t_{\alpha/2,n-1}) = \mathbb{P}(\mu \in \bar{X}_n \pm t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}) = 1-\alpha$$

- $1-\alpha$ **confidence interval** for the mean of a **normal population**: $\bar{x}_n \pm t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}$
- One-sided intervals: e.g., $1-\alpha$ **lower confidence bound** $\bar{x}_n - t_{\alpha,n-1}\frac{s_n}{\sqrt{n}}$

Comparing variances and F

- Compare the variances of two normal populations $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$
- Random samples $\{X_1, \cdots, X_n\}$ from $N(\mu_1, \sigma_1^2)$, $\{Y_1, \cdots, Y_m\}$ from $N(\mu_2, \sigma_2^2)$. Distribution of S_1^2/S_2^2?
- Let $U \sim \chi^2_{\nu_1}$, $V \sim \chi^2_{\nu_2}$ be independent. Then

$$F = \frac{U/\nu_1}{V/\nu_2}$$

  has an **F distribution** with ν_1 numerator df and ν_2 denominator df, denoted by F_{ν_1,ν_2}
- Suppose $T \sim t_\nu$. What is the distribution of T^2? Suppose $F \sim F_{\nu_1,\nu_2}$. What is the distribution of $1/F$?

CI for σ_1^2/σ_2^2

- Recall that $(n-1)S_1^2/\sigma_1^2 \sim \chi^2_{n-1}$, $(m-1)S_2^2/\sigma_2^2 \sim \chi^2_{m-1}$

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n-1,m-1}$$

- Let F_{α,ν_1,ν_2} be the $1-\alpha$ quantile of F_{ν_1,ν_2}

$$\mathbb{P}(F_{1-\alpha/2,n-1,m-1} \leq \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \leq F_{\alpha/2,n-1,m-1}) = 1-\alpha$$

- $1-\alpha$ confidence interval for σ_1^2/σ_2^2

$$\Big[\frac{s_1^2/s_2^2}{F_{\alpha/2,n-1,m-1}}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2,n-1,m-1}}\Big]$$

Constructing CI's in general

- Given a random sample $\{X_1, \cdots, X_n\}$, to construct a CI for a parameter θ of the unknown distribution
- Start with an appropriate statistic $h(X_1, \cdots, X_n, \theta)$
- From the sampling distribution of $h(X_1, \cdots, X_n, \theta)$, find a, b such that

$$\mathbb{P}(a \leq h(X_1, \cdots, X_n, \theta) \leq b) = 1 - \alpha$$

- Find the corresponding $L(X_1, \cdots, X_n), U(X_1, \cdots, X_n)$ such that

$$\mathbb{P}(L(X_1, \cdots, X_n) \leq \theta \leq U(X_1, \cdots, X_n)) = 1 - \alpha$$

- $1 - \alpha$ CI for θ: $[L(x_1, \cdots, x_n), U(x_1, \cdots, x_n)]$

- One-sided CI's can be constructed similarly
- We discussed constructing CI's for: population mean with large samples (z-intervals); normal population mean (t-intervals); normal population variance (using chi-squared distribution); ratio of variances of two normal populations (using F distribution)
- Other useful CI's: population proportion (z-intervals); difference of two population means (z- or t-intervals); difference of two population proportions (z-intervals)

6. Hypothesis testing

Ruppert 2011, Appendix A18

Example 0.1

A filling machine is designed to fill each bottle with 16 oz of beer. 20 samples give an average of 15.9 and a standard deviation of 0.2. Assume normality. Is there strong evidence that the true mean filling amount is different from 16 (and hence a recalibration of the machine, which will be costly)?

- Is a claim supported by data?
- **Null hypothesis** H_0: $\mu = 16$
- **Alternative hypothesis** H_1: $\mu \neq 16$
- H_0 is rejected only when there is strong evidence from data in support of H_1

A t-test

- Reject H_0 if sample mean is "significantly" different from $\mu_0 = 16$
- **Test statistic**

$$T_0 = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

- Given small $\alpha \in (0, 1)$ (e.g., 1%, 5%). If H_0 were true, $T_0 \sim t_{n-1}$, $|T_0| > t_{\alpha/2,n-1}$ has probability α, an unlikely event under H_0
- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$ (t_0 is the numerical value of T_0)
- This is a t-test

α, β, power

- α: **significance level**, is the probability of **type I error** (rejecting a true H_0)
- **Type II error** probability β: the probability of failing to reject a false H_0
- **Power** of a test $1 - \beta$: the probability of successfully rejecting a false H_0
- One controls type I error by selecting a small α, increases power of the test by using an appropriate sample size

Testing using CI's and p-values

- In Example 0.1, one rejects H_0 if μ_0 is not in the $1-\alpha$ confidence interval for the true mean

$$\mu_0 \notin \bar{x}_n \pm t_{\alpha/2,n-1} \frac{s_n}{\sqrt{n}}$$

- Alternatively, one could find the total area outside of $|t_0|$ under the t_{n-1} density (**p-value**) and compare with α

$$\text{p-value} = \mathbb{P}(|X| > |t_0|),\ X \sim t_{n-1}$$

- Reject H_0 if p-value is smaller than the significance level α
- p-value measures extremeness of data: probability to observe a value for the test statistic that is at least as extreme as t_0 if H_0 were true
- A small p-value is strong evidence against H_0

t-test in R

- When using data directly, use t.test function; when using summarized data, use tsum.test function in the BSDA library

```
> library(BSDA)
> tsum.test(mean.x=15.9, s.x=0.2, n.x=20,mu=16, conf.level=0.95)

        One-sample t-Test

data:  Summarized x
t = -2.2361, df = 19, p-value = 0.03754
alternative hypothesis: true mean is not equal to 16
95 percent confidence interval:
 15.8064 15.9936
sample estimates:
mean of x
     15.9
```

- With $\alpha = 5\%$, $|t_0| > t_{0.025,19} = 2.093$: reject H_0
  Alternatively, 95% CI doesn't contain $\mu_0 = 16$: reject H_0
  Alternatively, p-value$= 0.0375 < \alpha$: reject H_0

One-sided t-tests

- Suppose the alternative hypothesis is $H_1 : \mu < \mu_0$
- $H_0 : \mu = \mu_0, H_1 : \mu < \mu_0$. $T_0 = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ defined as before. Under H_0

$$\mathbb{P}(T_0 < -t_{\alpha,n-1}) = \mathbb{P}(\bar{X}_n < \mu_0 - t_{\alpha,n-1}\frac{S_n}{\sqrt{n}}) = \alpha$$

  Reject H_0 in support of H_1 if $\bar{x}_n < \mu_0 - t_{\alpha,n-1}\frac{s_n}{\sqrt{n}}$
- Alternatively, reject H_0 if $\mu_0 \notin (-\infty, \bar{x}_n + t_{\alpha,n-1}\frac{s_n}{\sqrt{n}}]$, a one-sided confidence interval for μ
- Alternatively, reject H_0 if p-value is less than α

$$\text{p-value} = \mathbb{P}(X < t_0), \ X \sim t_{n-1}$$

General procedure of hypothesis testing

- Determine H_1 (to show with strong evidence) and H_0
- Construct an appropriate test statistic
- Determine the sampling distribution of the test statistic under H_0
- For a given significance level α, determine when to (not to) reject H_0 in support of H_1
- We discussed testing the mean of a normal population
- Other useful tests: testing proportion, variance, difference in two means or proportions, ratios of two variances

Testing normality

- Is there strong evidence for non-normality? H_0: the sample comes from a normal distribution; H_1: not normal
- **Shapiro-Wilk** test: if H_0 were true, sample quantiles and normal quantiles should exhibit strong positive linear relation; a test statistic $W(X_1, \cdots, X_n)$ was constructed by Shapiro and Wilk; it should be close to 1 under H_0; reject H_0 otherwise
- Reject H_0 if p-value is small: *shapiro.test(index$rDJ)* returns a p-value of 2×10^{-11}. There is strong statistical evidence that the return of DJIA is not normal

Statistical vs practical significance

- **Statistical significance $\neq$ practical significance**
- E.g., when testing whether the mean filling amount $\mu \neq 16$: suppose you test 1000 bottles and obtain the following confidence interval for the mean: $[15.998, 15.999]$
- $H_0 : \mu = 16$ will be rejected since $16 \notin [15.998, 15.999]$; deviation from H_0 is statistically significant
- But practically, such a slight deviation might not be significant at all!
- When testing normality, it is helpful to use both normal plots and tests

7. Maximum likelihood estimation

Ruppert 2011, Chapter 5

Maximum likelihood estimation

- Point/interval estimation of means/variances
- Estimating unknown parameters (not necessarily mean/variance) of a certain distribution in general
- **Maximum likelihood estimation**: determine the values of the parameters so that the observed data are mostly likely from the distribution with these values
- Maximize the likelihood function

Example: discrete distribution

- Determine the success rate p of a Bernoulli distribution
- $X \sim$ Bernoulli distribution, $\mathbb{P}(X=1)=p, \mathbb{P}(X=0)=1-p$,

$$\mathbb{P}(X=x|p)=p^x(1-p)^{1-x}, \quad x=0,1$$

- Given observations $\{x_1,\cdots,x_n\}$. Probability (likelihood) of observing x_i is

$$p^{x_i}(1-p)^{1-x_i}$$

- **Likelihood function**

$$L(p)=\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}=p^{n\bar{x}_n}(1-p)^{n(1-\bar{x}_n)}, \quad \bar{x}_n=\frac{1}{n}\sum_{i=1}^n x_i$$

- To maximize $L(p)$, it is equivalent to maximize the **log likelihood function**

$$\log(L(p)) = n\bar{x}_n \log(p) + n(1 - \bar{x}_n) \log(1 - p)$$

Solving $d \log(L(p))/dp = 0$: $\bar{x}_n$ maximizes the likelihood function

- Given a random sample $\{X_1, \cdots, X_n\}$, the **maximum likelihood estimator** for p is $\hat{p} = \bar{X}_n$
- More generally, for a discrete distribution with pmf $\mathbb{P}(X = x|\theta) = p(x|\theta)$ or a continuous distribution with pdf $p(x|\theta)$, the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

Example: continuous distribution

- **Example**: Determine the maximum likelihood estimator for the parameter μ of an $N(\mu, 1)$ distribution.

  Log likelihood function

$$
\begin{aligned}
\log(L(\mu)) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i-\mu)^2}{2})\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}
$$

  Solving $d\log(L(\mu))/\mu = \sum_{i=1}^{n}(x_i-\mu) = 0$: the MLE for μ is $\hat{\mu} = \bar{X}_n$

- **Example**: Determine the MLEs for (μ, σ^2) of $N(\mu, \sigma^2)$

$$\begin{aligned}\log(L(\mu,\sigma^2)) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i-\mu)^2}{2\sigma^2})\right) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

Solving

$$\frac{\partial \log(L(\mu,\sigma^2))}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\frac{\partial \log(L(\mu,\sigma^2))}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

The MLEs are $\hat{\mu} = \bar{X}_n, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Accuracy and precision of MLEs

- The MLE $\hat{\sigma}^2$ is biased: the bias $\mathbb{E}[\hat{\sigma}^2] - \sigma^2 = -\frac{1}{n}\sigma^2$ converges to 0 as $n \to +\infty$
- Let $\hat{\theta}$ denote the MLE of θ. Under mild conditions, as $n \to +\infty$
  - **Consistency**: $\hat{\theta} \to \theta$ in probability
  - **Asymptotic normality**

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})} \to N(0, 1)$$

  - **Efficiency**: compared to other estimators, when sample size is large, MLE has the smallest variance (nearly MVUE)
- **Invariance**: if $\hat{\theta}$ is the MLE of θ, $f(\hat{\theta})$ is the MLE of $f(\theta)$

SE of MLE

- **Fisher information**

$$I_n(\theta) = -\mathbb{E}\Big[\frac{d^2}{d\theta^2}\log(\prod_{i=1}^{n} p(X_i|\theta))\Big]$$

- Standard error of MLE

$$se(\hat{\theta}) \approx \sqrt{\frac{1}{I_n(\theta)}}$$

Estimated standard error $\sqrt{1/I_n(\hat{\theta})}$

- This allows one to construct confidence intervals

Example: Bernoulli distribution

- Recall the log likelihood function

$$\log(L(p)) = n\bar{x}_n \log(p) + n(1-\bar{x}_n)\log(1-p)$$

$$I_n(p) = -\mathbb{E}\Big[-\frac{n\bar{X}_n}{p^2} - \frac{n(1-\bar{X}_n)}{(1-p)^2}\Big] = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}$$

- Standard error of the MLE $\hat{p} = \bar{X}_n$ is thus

$$\sqrt{\frac{p(1-p)}{n}}$$

- Consistent with a direct calculation (compute the standard deviation of $\bar{X}_n$)

8. Likelihood ratio tests

Ruppert 2011, Chapter 5

Likelihood ratio tests

- Learned how to test mean/variance of a normal distribution
- What if distribution is more general?
- Parameter of interest is arbitrary?
- No obvious test statistic to use?
- Null and alternative hypotheses more complex?
- Testing multiple parameters: $H_0 : (\mu, \sigma) = (0, 1)$ vs $H_1 : (\mu, \sigma) \neq (0, 1)$
- $H_0 : \mu \in (0, 1)$ vs $H_1 : \mu \notin (0, 1)$
- Likelihood ratio tests (LRTs) are widely used in these situations

Principles of LRT

- Construct test statistic using maximum likelihood estimator(s)
- To test $\theta \in \mathbb{R}^m$ with $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ ($\theta_0 \in \mathbb{R}^m$)
- Let $\hat{\theta}_n$ be the maximum likelihood estimator, and $\log(L(\hat{\theta}_n))$ the log likelihood function at $\hat{\theta}_n$

$$\log(L(\hat{\theta}_n)) = \max_{\theta} \log(L(\theta))$$

- Let $\log(L(\theta_0))$ be the log likelihood function at the hypothesized value θ_0

$$\log(L(\hat{\theta}_n)) \geq \log(L(\theta_0))$$

- If H_0 were true, $\log(L(\hat{\theta}_n))$ and $\log(L(\theta_0))$ are close
- Reject H_0 if the following is large

$$2[\log(L(\hat{\theta}_n)) - \log(L(\theta_0))] = 2\log\frac{L(\hat{\theta}_n)}{L(\theta_0)}$$

- Under $H_0 : \theta = \theta_0 \in \mathbb{R}^m$, the above approximately follows χ^2_m
- At significant level $\alpha \in (0, 1)$, reject H_0 if

$$2\log\frac{L(\hat{\theta}_n)}{L(\theta_0)} > \chi^2_{\alpha,m}$$

where $\chi^2_{\alpha,m}$ is the $1-\alpha$ quantile of χ^2_m

LRT Example

- Given a sample of size 100 from a normal population. $\hat{\mu} = \bar{x}_n = 0.21$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = 2.33$. Are the data significantly different from $N(0,1)$? Significance level $= 5\%$.

  $H_0 : \mu = \mu_0 = 0, \sigma = \sigma_0 = 1;\ \chi^2_{0.05,2} = 5.99$

$$
\begin{aligned}
\log(L(\hat{\mu}, \hat{\sigma}^2)) &= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \\
&= -\frac{n}{2}(\log(2\pi\hat{\sigma}^2) + 1) = -184.19
\end{aligned}
$$

- Can verify that $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + n\bar{x}_n^2$

$$\begin{aligned}
\log(L(\mu_0, \sigma_0^2)) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} x_i^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}(\hat{\sigma}^2 + \bar{x}_n^2) = -210.60
\end{aligned}$$

Reject H_0 since $2(-184.19 + 210.60) = 52.82 > 5.99$

9. Resampling

Ruppert 2011, Chapter 6

Sampling distributions

- When estimating an unknown mean μ, use sample mean $\bar{X}_n$
  - For normal population, $(\bar{X}_n - \mu)/(S_n/\sqrt{n})$ is t_{n-1}
  - Otherwise, for a large sample, the above is approximately normal
  - Important to know sampling distributions
  - so that we can construct confidence intervals for μ, test hypotheses on μ
- Need to know sampling distributions for more general estimators: e.g., sample median for estimating an unknown median; maximum likelihood estimators, etc.

- For a **maximum likelihood estimator**, one may use its asymptotic normality and **Fisher information**
- In general, computing standard error of an arbitrary estimator and constructing confidence intervals can be difficult due to lack of explicit sampling distributions: e.g., sample median
- **Resampling** (or **bootstrap**): use simulation to approximate standard errors and construct confidence intervals numerically
- Let $T = g(X_1, \cdots, X_n)$ be a statistic, where $X_i \sim F$, F is unknown
- Two steps to approximate $\mathrm{var}_F(T)$

  1. Approximate F by the empirical cdf $\hat{F}$, $\mathrm{var}_F(T)$ by $\mathrm{var}_{\hat{F}}(T)$
  2. Use **simulation** to estimate $\mathrm{var}_{\hat{F}}(T)$

Empirical cdf

- For a random sample $\{X_1, \cdots, X_n\}$, **empirical cdf** is given by

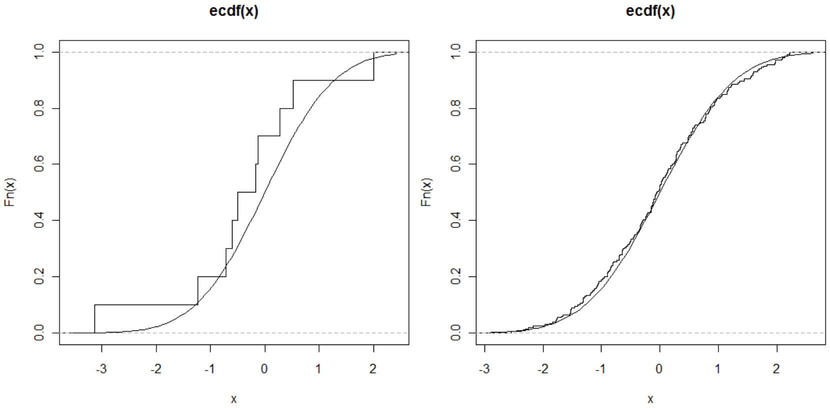
$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$$

  i.e., the fraction of observations that are $\leq x$
- As n increases, $\hat{F}(x)$ converges to $F(x)$ (LLN)
- Empirical cdf for data from $N(0, 1)$

```
x=rnorm(200)
Fhat=ecdf(x)
plot(Fhat,verticals=T,do.points=F)
curve(pnorm(x),add=T)
```

Empirical cdf for data from $N(0,1)$

- Sample size 10 and 200 respectively



Simulate from empirical cdf

- $\hat{F}$ is a **step function** with step size $1/n$ at X_i: it is the cdf of a discrete distribution with probability $1/n$ at each X_i
- **Resampling**: to simulate a sample of size n from $\hat{F}$, randomly pick a number from $\{X_1, \cdots, X_n\}$, repeat for n times, one obtains
$$\{X_1^*, \cdots, X_n^*\}$$
- Compute $T^* = g(X_1^*, \cdots, X_n^*)$. Repeat for B times to obtain $T_1^*, \cdots, T_B^*$. Estimate $\mathrm{var}_{\hat{F}}(T)$ by
$$\frac{1}{B-1}\sum_{b=1}^{B}(T_b^* - \bar{T}^*)^2, \bar{T}^* = \frac{1}{B}\sum_{b=1}^{B} T_b^*$$

Example: standard error of sample median

- Given Amazon's daily return from 4/26/2011 to 9/15/2011 ($n = 100$ observations)

```
amazon = read.csv("amazon.csv",header=T)
x=amazon$rAmazon[1:100]
median(x)
n=length(x)
nbootstrap=5000
resample=rep(0,nbootstrap)
for (b in 1:nbootstrap)
{
xstar=sample(x,n,replace=T)
resample[b]=median(xstar)
}
sd(resample)
```

- Sample median (of 100 returns): 0.0032; standard error of the sample median: 0.0029

Standard error of sample mean

- Repeat the above with sample mean replacing sample median

  sample mean (of 100 returns): 0.0020
  standard error computed using bootstrap: 0.00240

- Alternatively, one can compute the standard error using

$$\frac{s}{\sqrt{n}} = 0.00242, \text{ a nice match!}$$

Here s is the sample standard deviation of the $n = 100$ returns

Confidence intervals

- Let θ be an unknown parameter of a certain distribution
- Given data $\{X_1, \cdots, X_n\}$, how to construct a confidence interval for θ
- Let $\hat{\theta} = g(X_1, \cdots, X_n)$ be an estimator for θ; in general one does not know the distribution of $\hat{\theta}$
- Use bootstrap to approximate the distribution

- Distribution of $\hat{\theta}$ depends on F, the unknown population. Construct $\hat{F}$ from $\{X_1, \cdots, X_n\}$
    1. Approximate F by $\hat{F}$, distribution of $\hat{\theta} = g(X_1, \cdots, X_n)$ by that of $\hat{\theta}^* = g(X_1^*, \cdots, X_n^*)$, $X_i^* \sim \hat{F}$ and i.i.d
    2. Approximate distribution of $\hat{\theta}^*$ by simulation
- Simulate $\hat{\theta}_b^*, 1 \leq b \leq B$. One can then estimate
    1. $\mathbb{P}(\hat{\theta}^* \leq x)$ which approximates $\mathbb{P}(\hat{\theta} \leq x)$
    2. $\text{var}_{\hat{F}}(\hat{\theta}^*)$ which approximates $\text{var}_F(\hat{\theta})$
    3. $\mathbb{E}_{\hat{F}}[\hat{\theta}^*]$ which approximates $\mathbb{E}_F[\hat{\theta}]$, etc.

- Given $\alpha \in (0, 1)$. For a confidence interval of the form $\theta \in (\hat{\theta} + a, \hat{\theta} + b)$, want to find a, b such that

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(\hat{\theta} + a \leq \theta \leq \hat{\theta} + b) \\ &= \mathbb{P}(-b + \theta \leq \hat{\theta} \leq -a + \theta) \\ &\approx \mathbb{P}(-b + \theta \leq \hat{\theta}^* \leq -a + \theta) \end{aligned}$$

- Let q_L^* be the $\alpha/2$ sample quantile of $\{\hat{\theta}_b^*, 1 \leq b \leq B\}$, q_U^* be the $1 - \alpha/2$ sample quantile. It suffices that

$$-a + \theta \approx q_U^*, \quad -b + \theta \approx q_L^*$$

Approximate θ by $\hat{\theta}$, we obtain $a \approx \hat{\theta} - q_U^*$, $b \approx \hat{\theta} - q_L^*$ and an approximate CI $(2\hat{\theta} - q_U^*, 2\hat{\theta} - q_L^*)$

CI for the median

- With $\alpha = 5\%$, *quantile(resample,probs=c(0.025,0.975))* computes the sample quantiles of the sample median

$$q_L^* = -0.0053, \quad q_U^* = 0.0077$$

Approximate CI for the median daily return of Amazon's stock:

$$(2\hat{\theta} - q_U^*, 2\hat{\theta} - q_L^*) = (-0.0012, 0.0118)$$

where $\hat{\theta} = 0.0032$ is the sample median of the 100 returns

CI for the mean

- With $\alpha = 5\%$, the sample quantiles of the sample mean

$$q_L^* = -0.0028, \quad q_U^* = 0.0069$$

Approximate CI for the mean daily return of Amazon's stock:

$$(2\hat{\theta} - q_U^*, 2\hat{\theta} - q_L^*) = (-0.0028, 0.0068)$$

where $\hat{\theta} = 0.0020$ is the sample mean of the 100 returns

- Alternatively, use the CLT to construct a CI:
  $\bar{x}_n \pm z_{\alpha/2} s/\sqrt{n} = (-0.0027, 0.0068)$, a nice match!

CLT for sample median

- **Theorem**. Let θ be the median of a distribution with a continuous density $f(x)$. Then for a large enough sample size n, the sample median $\hat{\theta}$ is approximately normal with mean θ and variance
$$(se(\hat{\theta}))^2 = \frac{1}{4n(f(\theta))^2}$$
- Note the characteristic $1/\sqrt{n}$ convergence
- Not directly applicable due to the lack of the density
- But the variance can be computed using resampling as previously described

CI for the median using CLT

- By the CLT for the sample median $\hat{\theta}$,

$$1 - \alpha \approx \mathbb{P}\Big(- z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{se(\hat{\theta})} \leq z_{\alpha/2} \Big) = \mathbb{P}(\theta \in \hat{\theta} \pm z_{\alpha/2} se(\hat{\theta}))$$

- We obtain a CI for the median that is symmetric around the sample median: $\hat{\theta} \pm z_{\alpha/2} se(\hat{\theta})$
- CI's constructed using the above CLT is sensitive to sample size

- Use Amazon's prices from 5/16/1997 to 9/15/2011 (3600 prices and 3599 returns)

	$n = 100$	$n = 3599$
mean	0.0020	0.0014
SE classical	0.00242	0.00073
SE resampling	0.00240	0.00072
CI CLT	(-0.0027,0.0068)	(-0.00008,0.0028)
CI resampling	(-0.0028,0.0068)	(-0.00006,0.0027)
median	0.0032	0
SE	0.0029	0.00045
CI resampling	(-0.0012,0.0118)	(-0.0006,0.0012)
CI CLT	(-0.0027,0.0091)	(-0.0009,0.0009)

10. Modeling univariate variables

Ruppert 2011, Chapter 5

- **Parsimonious**: models that fit data well but without too many parameters

  Too few parameters: cannot capture main features of data
  Too many parameters: more estimation error/overfitting

- **Parametric** model: assume a specific distribution, with unknown parameters
- Financial data exhibit distributions with skewness and heavy (fat) tails: extreme movements are common
- Skewness of any symmetric distribution is 0, kurtosis of a normal distribution is 3
- Want distributions with desired skewness and kurtosis

Heavy tails from mixture

- Let $Z \sim N(0,1)$, $U \sim U(0,1)$ be independent of Z and

$$Y = \begin{cases} Z, & U < 0.9 \\ 5Z, & U \geq 0.9 \end{cases} = \sqrt{V}Z, \quad V = \begin{cases} 1, & U < 0.9 \\ 25, & U \geq 0.9 \end{cases}$$

  90% of time in a low variance regime, and 10% of time in a high variance regime. The cdf of Y is

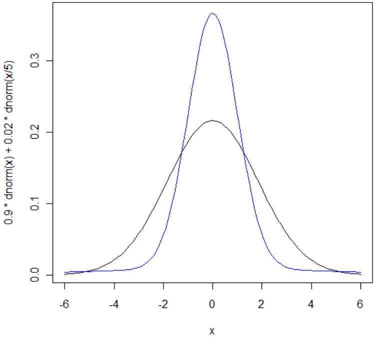
$$\mathbb{P}(Y \leq y) = \mathbb{P}(Y \leq y | U < 0.9)\mathbb{P}(U < 0.9)$$

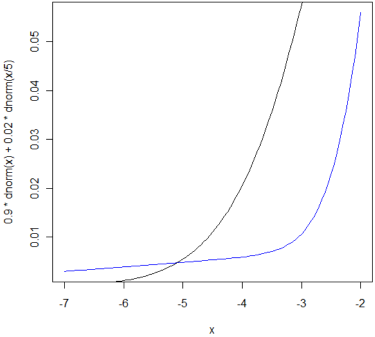
$$+\mathbb{P}(Y \leq y | U \geq 0.9)\mathbb{P}(U \geq 0.9) = 0.9N(y) + 0.1N(y/5)$$

- Mean 0, variance 3.4, skewness 0, kurtosis 16.45

Normal mixture

- curve(0.9*dnorm(x)+0.02*dnorm(x/5),from=-6,to=6,col="blue")
  curve(dnorm(x,mean=0,sd=sqrt(3.4)),from=-6,to=6,add=T)





- Compare $\mathbb{P}(Y < -6)$ and $\mathbb{P}(N(0, 3.4) < -6)$

$$\mathbb{P}(N(0, 3.4) < -6) = N(-6/\sqrt{3.4}) = N(-3.25) = 0.00057$$

$$\mathbb{P}(Y < -6) = 0.9N(-6) + 0.1N(-1.2) = 0.0115$$

  20 time more likely for Y to drop below -6
- Large observations are much more likely in a normal mixture model
- More **general normal mixture**: $Y = \mu + \sqrt{V}Z$ for a positive r.v. V
- t-distribution is such a mixture: $T \sim t_\nu, X \sim \chi^2_\nu, Z \sim N(0, 1)$

$$T = \frac{Z}{\sqrt{X/\nu}} = \sqrt{V}Z, \ \ V = \nu/X$$

Normal inverse Gaussian model

- The **normal inverse Gaussian** (NIG) model is a popular Lévy process model allowing jumps. Asset price at time t:

$$S_t = S_0 e^{X_t}$$

- **Log return** over the period $[0, t]$: $X_t \sim \mu t + \beta z_t + \sqrt{z_t} Z$, where $Z \sim N(0,1)$, $z_t \sim IG(\delta t, \gamma)$ has an **inverse Gaussian distribution**
- Density of $IG(a, b)$

$$p(x) = \frac{a}{\sqrt{2\pi x^3}} \exp\left(- \frac{(a - bx)^2}{2x} \right), x > 0$$

Density of NIG process

- Modified Bessel function of the second kind with order 1

$$K_1(x) = \int_0^{\infty} \cosh(y) e^{-x\cosh(y)} dy, \quad \cosh(y) = (e^y + e^{-y})/2$$

- The pdf of X_t is then

$$p_t(x) = \frac{\alpha\delta t}{\pi} \frac{K_1(\alpha\sqrt{\delta^2 t^2 + (x-\mu t)^2})}{\sqrt{\delta^2 t^2 + (x-\mu t)^2}} \exp(\delta\gamma t + \beta(x - \mu t))$$

where $\alpha > 0, -\alpha \leq \beta \leq \alpha, \delta > 0, \mu \in \mathbb{R}$, $\gamma = \sqrt{\alpha^2 - \beta^2}$

Characteristic function of NIG process

- Transform methods for options valuation: characteristic function of X_t

$$\phi_t(\xi) = \exp\left(i\mu t\xi - \delta t\left(\sqrt{\alpha^2 - (\beta + i\xi)^2} - \sqrt{\alpha^2 - \beta^2}\right)\right)$$

- Nonzero β introduces skewness; any desired kurtosis

$$\text{skewness} = \frac{3\beta}{\alpha\sqrt{\delta t\gamma}}$$

$$\text{kurtosis} = \frac{3(\alpha^2 + 4\beta^2)}{\alpha^2\delta t\gamma}$$

Fitting NIG and BSM

- In the Black-Scholes-Merton model

$$S_t = S_0 e^{X_t}, \quad X_t = \mu t + \sigma B_t \sim N(\mu t, \sigma^2 t)$$

  Here $\{B_t, t \geq 0\}$ is a standard Brownian motion
- Given daily data, assuming 252 trading days per year, $t = 1/252$
- **BSM**: daily log return X_t has normal density with unknown parameters μ, σ
- **NIG**: daily log return X_t has density given previously with unknown parameters $\alpha, \beta, \mu, \delta$
- Find maximum likelihood estimates for the parameters; compare the models

Fitting Amazon's return

- Amazon's daily prices in 2013

	Date	price	return
1	12/31/2013	398.79	0.01368432
2	12/30/2013	393.37	-0.01190235
...			
251	1/3/2013	258.48	0.00453674
252	1/2/2013	257.31	NA

- An optimization problem needs be solved to obtain MLEs

Optimization in R

- Maximum likelihood estimation

$$\min_{lowerbound \leq \mu, \sigma, \nu, \xi \leq upperbound} \left(-\sum_{i=1}^{n} \log(density(x_i, parameters)) \right)$$

  where density is the proposed density
- The minimization problem solved numerically
- R function for minimization:
  optim(initialvalue,function,method,lowerbound,upperbound)

```
> #fitting NIG
  library(fGarch)
  amazon=read.csv("amazon2013.csv", header=T)
  x=amazon$return[1:nrow(amazon)-1]
  t=1/252;

  #NIG Fitting, theta[1]=alpha, theta[2]=delta, theta[3]=mu, theta[4]=beta
  initialvalue=c(50,2,0,0);
  nig=function(x,theta){theta[1]*theta[2]*t/pi*besselK(theta[1]*sqrt(theta[2]^2*t^2+(x-theta[3]
  *t)^2),1)/(sqrt(theta[2]^2*t^2+(x-theta[3]*t)^2))*exp(theta[2]*sqrt(theta[1]^2-theta[4]^2)*t
  +theta[4]*(x-theta[3]*t))}
  result=optim(initialvalue,fn=function(theta){-sum(log(nig(x,theta)))},
               method="L-BFGS-B")
  result
```

- Output from R for the NIG fitting:
  $par
  [1] *81.09801849 5.51334237 -0.07008602 7.46965454*
  $value
  [1] *-680.5536*
- Optimization is successful, and returns

  $$\alpha = 81.098018, \delta = 5.513342, \mu = -0.070086, \beta = 7.469655$$

  with log likelihood 680.5536
- For the BSM fitting

  $$\mu = 0.439899, \sigma = 0.2680215$$

  with log likelihood 668.2785

- Given data, compare models using the log likelihood function

$$\log(L(\hat{\theta}))$$

  When the log likelihood is larger, the model is more consistent with the data

- **Akaike's information criterion** (pronunciation: roughly ah-kah-ee-kay): take the number of parameters (denoted by n) into account to avoid overfitting

$$-2(\log(L(\hat{\theta})) - n)$$

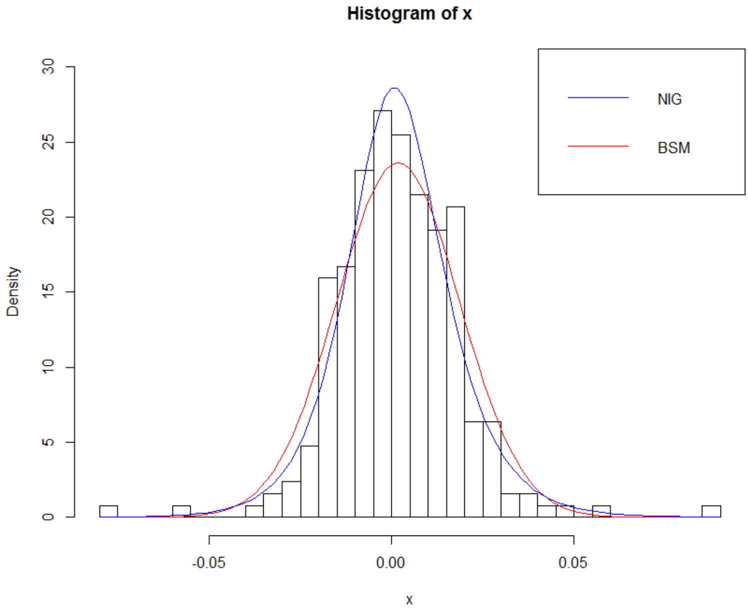
  The smaller the AIC, the better the model

Model selection using AIC

- Comparing AIC

distribution	# of parameters	AIC
NIG	4	$-2(680.5536 - 4) = -1353.107$
BSM	2	$-2(668.2785 - 2) = -1332.557$

- NIG provides a better fit


Histogram of x
NIG
BSM
Density
30
25
20
15
10
5
0
-0.05
0.00
0.05
x

10. Modeling multivariate data

Ruppert 2011 Chapter 7

Modeling multivariate distributions

- Interested in joint behavior of several financial variables
- Probabilistic modeling of multiple random variables
- Commonly used multivariate distributions
- Parameter estimation

Basics

- Suppose ith financial variable is denoted by Y_i, $1 \leq i \leq d$

$$Y = (Y_1, \cdots, Y_d)^\top$$

- $\mathbb{E}[Y]$ is a d-vector: $\mathbb{E}[Y] = (\mathbb{E}[Y_1], \cdots, \mathbb{E}[Y_d])^\top$
- Covariance matrix describes linear relations among the variables

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{dd} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ & \vdots & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix}$$

where $\sigma_{ij} = \sigma_{ji} = \mathrm{cov}(Y_i, Y_j)$, $\sigma_{ii} = \sigma_i^2 = \mathrm{var}(Y_i)$; Σ is symmetric and positive semi-definite

Correlation

- Correlation matrix more informative in describing linear relations

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{dd} \\ \rho_{21} & 1 & \cdots & \rho_{2d} \\ & \vdots & & \\ \rho_{d1} & \rho_{d2} & \cdots & 1 \end{pmatrix}$$

  where $\rho_{ij} = \rho_{ji} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \in [-1, 1]$; ρ is symmetric and positive semi-definite
- Standardization: $\text{cov}\left(\frac{Y_i - \mathbb{E}[Y_i]}{\sigma_i}, \frac{Y_j - \mathbb{E}[Y_j]}{\sigma_j}\right) = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \rho_{ij}$. After standardizing Y, the covariance matrix equals the correlation matrix

Basic rules

- Expectation is linear: $\mathbb{E}[a^\top Y] = a^\top \mathbb{E}[Y]$, $a \in \mathbb{R}^d$
- For $a, b \in \mathbb{R}^d$, $\text{cov}(a^\top Y, b^\top Y) = a^\top \Sigma b$. In particular, $\text{var}(a^\top Y) = a^\top \Sigma a$
- Markowitz's mean-variance portfolio theory
  Y_i: return of ith asset
  a_i: proportion invested in asset i

$$\min_a \text{var}(a^\top Y)$$

subject to

$$\mathbb{E}[a^\top Y] = r, \quad \sum_{i=1}^{d} a_i = 1, \quad etc.$$

Multivariate normal distributions

- pdf of a multivariate normal r.v. Y with mean μ and *positive definite* covariance matrix Σ

$$p(y) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right)$$

- Properties
  1. marginal distributions are normal
  2. linear combinations of Y_i's are normal
  3. $\rho_{ij} = 0 \Rightarrow Y_i, Y_j$ are independent

Multivariate t distributions

- Multivariate normal distributions: not heavy enough tails for financial variables
- A r.v. Y with multivariate t distribution exhibits heavier tails, and admits similar properties: marginal distributions are still t; a linear combination of Y_i's is still t
- Recall the following representation for t_ν

$$\frac{Z}{\sqrt{X/\nu}} \sim t_\nu, \quad Z \sim N(0,1) \text{ and } X \sim \chi^2_\nu \text{ independent}$$

- For $\mu \in \mathbb{R}^d, \Lambda \in \mathbb{R}^d \times \mathbb{R}^d$, let

$$Y = \mu + \frac{Z}{\sqrt{X/\nu}}, \quad Z \sim N(0, \Lambda) \text{ and } X \sim \chi^2_\nu \text{ independent}$$

- When $\nu > 1$, $\mathbb{E}[Y] = \mu$; when $\nu > 2$, the covariance matrix of Y is $\Sigma = \frac{\nu}{\nu-2}\Lambda$
- For a multivariate t r.v. Y, **zero correlation does not imply independence** due to the common denominator $\sqrt{X/\nu}$

Tail dependence

- Multivariate t distributions exhibit **tail dependence**: extreme observations tend to occur together (if X is near zero, all components of Y will be extreme)
- Y_1 and Y_2 exhibit lower tail dependence if the **coefficient of lower tail dependence**

$$\lambda_l = \lim_{q \downarrow 0} \mathbb{P}(Y_2 \leq F_{Y_2}^{-1}(q) | Y_1 \leq F_{Y_1}^{-1}(q)) > 0$$

where $F_{Y_i}^{-1}(q)$ is the q quantile of Y_i; i.e., if Y_1 is extreme, Y_2 tends to be extreme as well; **coefficient of upper tail dependence**

$$\lambda_u = \lim_{q \uparrow 1} \mathbb{P}(Y_2 \geq F_{Y_2}^{-1}(q) | Y_1 \geq F_{Y_1}^{-1}(q))$$

Correlation vs tail dependence

- If Y_1 and Y_2 are independent, $\lambda_l = \lambda_u = 0 \Rightarrow$ no tail dependence; if there is tail dependence, then Y_1, Y_2 are dependent (but not necessarily correlated!)
- **Correlation doesn't imply tail dependence**: multivariate normal distributions do not exhibit tail dependence
- **Tail dependence doesn't imply correlation**: multivariate t distributions exhibit tail dependence, even for zero correlation
- Financial variables often exhibit tail dependence, multivariate t could be attractive

Fisher's information and confidence interval

- For multiple parameters $\theta = (\theta_1, \cdots, \theta_n)$, the MLE $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_n)$ is **asymptotically normal**

$$\hat{\theta} \sim N(\theta, I_n^{-1}(\theta))$$

  where $I_n(\theta)$ is **Fisher's information** with ijth entry given by

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(L(\theta))\right]$$

- Approximate **confidence intervals** for θ_i can be constructed (when the sample size is large enough!)
- **Resampling** often used to find confidence intervals