Lecture 4: Univariate Descriptive Statistics/EDA

Brian Thelen
258 West Hall
bjthelen@umich.edu

Statistics 509 - Winter 2022
Ref: Ruppert/Matteson: Chapter 4.1-4.5

Overview of Lecture.

- Descriptive statistics
  - Summary quantitative statistics
  - Graphical summaries
    - histograms
    - density estimation
- Assessing probability distribution models
  - QQ Plots
  - Intro to goodness-of-fit tests

Background on R. Need to include:

```
> source('startup.R')
```

- Some new functions in `startup.R`

Some Summary Statistics - Review from Lecture 3

Background. Suppose that $X \sim F$ and have sample $x_1, x_2, \ldots, x_n$, and central moments of μ_k, m_k for $k = 2, 3, 4$.

Parameters/Statistics	Distn Parameter	Sample Statistic
Standard deviation	$\sigma = \sqrt{\mu_2}$	$\mathrm{SD}(x) = \sqrt{m_2}$
Skewness	$\frac{\mu_3}{(\mu_2)^{\frac{3}{2}}}$	$\frac{m_3}{(m_2)^{\frac{3}{2}}}$
(Excess) Kurtosis	$\frac{\mu_4}{\mu_2^2} - 3$	$\frac{m_4}{m_2^2} - 3$

Remarks.

- Skewness is a measure of the asymmetry of the distribution/sample values
- Kurtosis is a measure of how heavy-tailed the distribution/sample values

More on Skewness/Kurtosis

Normal Distribution

- For $X \sim \mathcal{N}(\mu, \sigma^2)$, the skewness and kurtosis are 0, 0
- For a random sample of $X_1, X_2, \ldots, X_n$ from $\sim \mathcal{N}(\mu, \sigma^2)$
  - the sample skewness and sample kurtosis should be relatively close to 0, 0
  - Expected "closeness" of the sample values depends on sample size – more as sample size increases
- There are statistical hypothesis tests for normality based on skewness and/or kurtosis

Double Exponential Distribution

- For $X \sim \text{DExp}(\mu, \lambda)$, skewness is 0 and kurtosis is 3
- For a random sample of $X_1, X_2, \ldots, X_n$ from $\text{DExp}(\mu, \lambda)$
  - the sample skewness and sample kurtosis should be relatively close to 0 and 3 , respectively
  - Expected "closeness" of these sample values depends on sample size – more as sample size increases

Examples - Skewness and Kurtosis.

Data	Skewness	Kurtosis
500 random deviates from $\mathcal{N}(0,1)$	0.0980	0.2011
500 random deviates from DExp$(0,1)$	0.3796	2.9360
500 random deviates from Exp(1)	1.6169	3.3536
2480 values – SP500 log(weekly returns) 1960-2007	-0.3662	3.3870

R-Session (Commands and Output)

```
library(fExtremes)  # Needed for skewness and kurtosis
xnorm <- rnorm(500,0,1)
> skewness(xnorm)
[1] 0.09803232
> kurtosis(xnorm)
[1] 0.2011059

> xdexp <- rdexp(500,0,1)
> skewness(xdexp)
[1] 0.3796158
> kurtosis(xdexp)
[1] 2.936092
```

```
> xexp <- rexp(500,1)
> skewness(xexp)
[1] 1.616941
> kurtosis(xexp)
[1] 3.353643
```

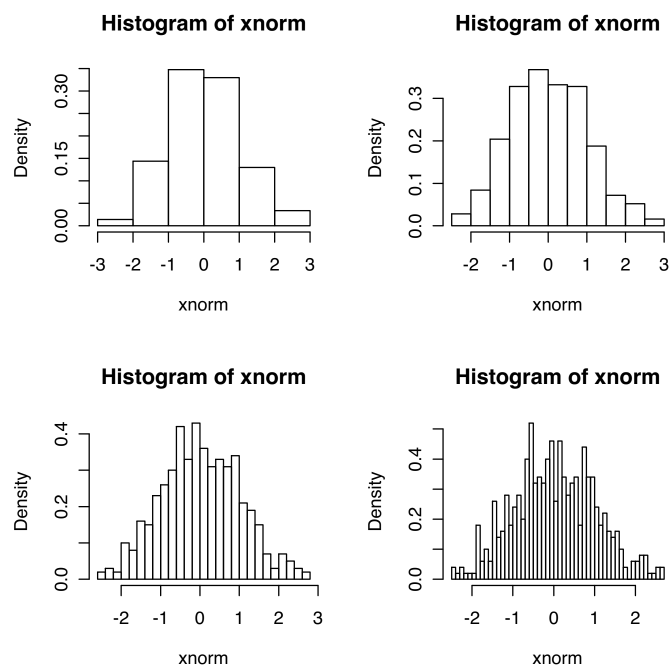
```
> X = read.csv("Data\\SP500_wkly_Jan1_60_Jul23_07.csv",header=TRUE)
> SP500wk <- rev(X\$Close)
> SP500wk_lreturn <- diff(log(SP500wk)) # log returns (weekly)
> skewness(SP500wk_lreturn)
[1] -0.3662413
> kurtosis(SP500wk_lreturn)
[1] 3.387008
```

Histograms

Background. Have sample $x_1, x_2, \ldots, x_n$ and want the histogram to be a good representation of the "distribution."

- Histograms – area of rectangles correspond to frequency

Examples: Below are histograms with # rectangles being 6, 11, 21, and 41 (R-variable "breaks"=5,10,20,40)



过度强调error

Question. Which one is preferred? break=10/20

R-code

```
xnorm <- rnorm(500,0,1)
par(mfrow=c(2,2)) # setting up for a 2 x 2 arrangement of subplots

hist(xnorm,breaks = 5,freq=FALSE)
hist(xnorm,breaks = 10,freq=FALSE)
hist(xnorm,breaks = 20,freq=FALSE)
hist(xnorm,breaks = 40,freq=FALSE)
```

Density Estimation

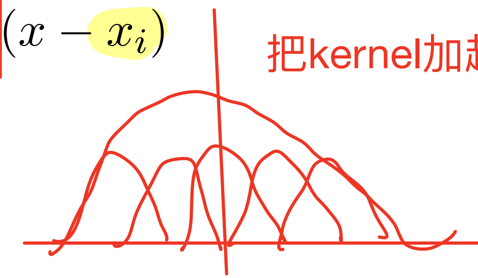
Remark. Histogram is a “coarse” (piecewise constant) density estimate – can do better.

Definition. For sample data $x_1, x_2, \ldots, x_n$, a kernel-based density estimate is defined as

kernel with bandwidth b

$$\hat{f}_b(x) = \frac{1}{n} \sum_{i=1}^{n} \boxed{K_b}(x - x_i)$$

把kernel加起来



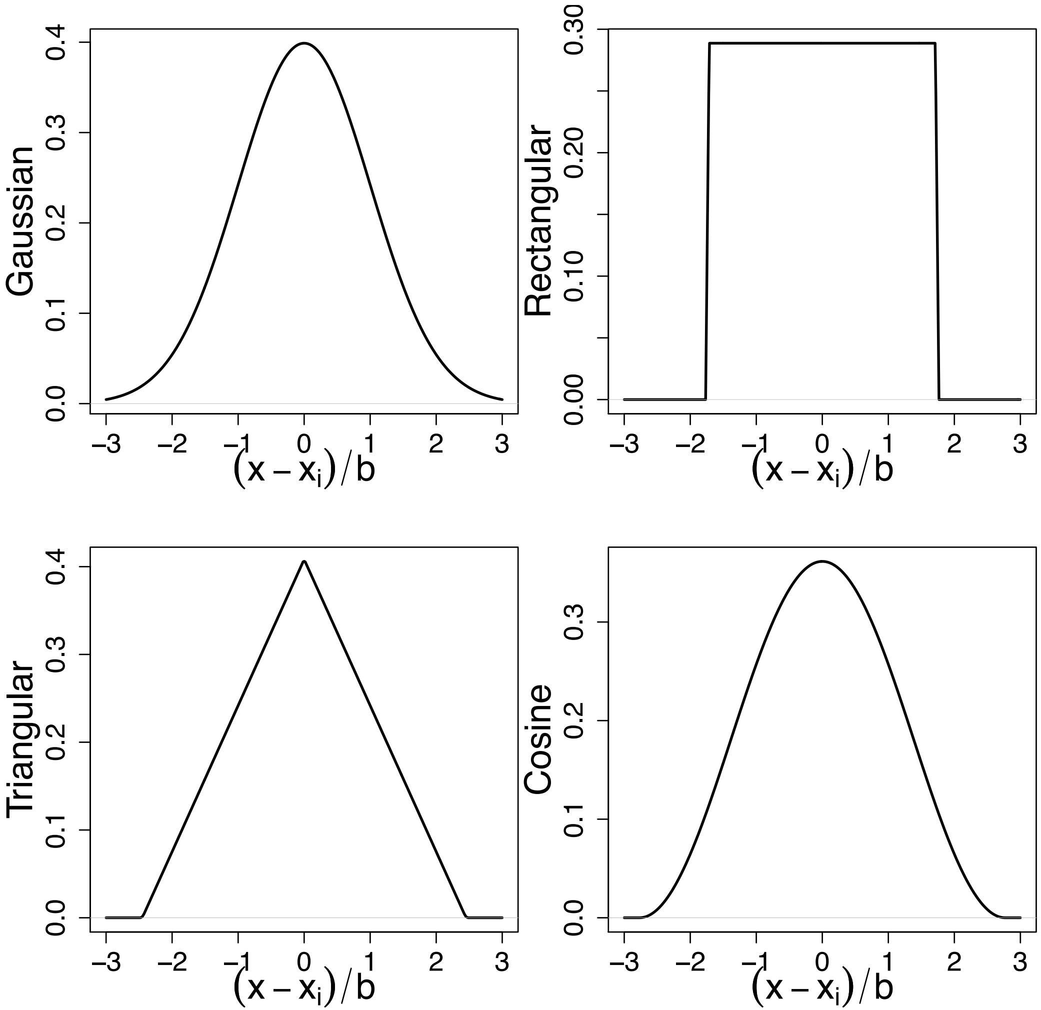
where

- $K_b(x - x_i) = \frac{1}{b} K\left(\frac{x - x_i}{b}\right)$
- K is called the **kernel function** – this function integrates to 1 and has a standard deviation of 1 K可以是任何density function
  - Possible shapes for K are "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" or "optcosine"

很难选择kernel和b

- In **R**, b is **bandwidth parameter** (positive number) and essentially is the standard deviation of K_b

Examples of $K((x - x_i)/b)$



More on Density Estimation

Remark. There are differing definitions of bandwidth parameter

Remark. Effect of bandwidth is

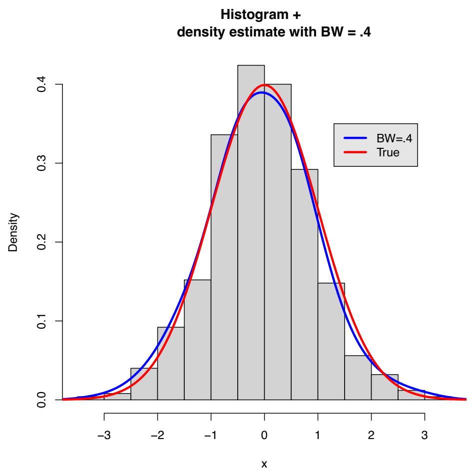
- When bw parameter b gets large,

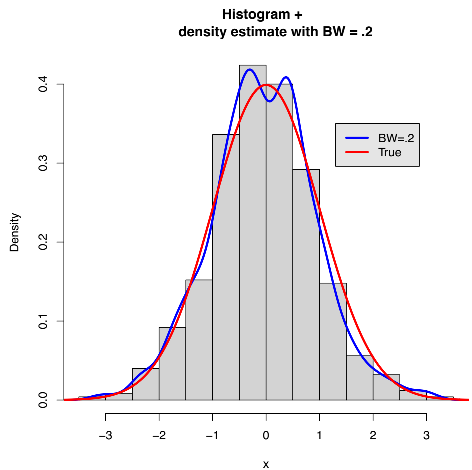
  K gets broader less peaked and more smoother pdf

- When bw parameter b gets small,

  K gets thinner, more peaked and less smoother
  会变得越来越敏感

Example





Question. What is the expected value of $\hat{f}_b(x)$ if have an iid sample from distribution with pdf f?

Answer. $E[\hat{f}_b(x)] = E[\frac{1}{n} \sum K_b(x - x_i)]$

$$= \frac{1}{n} \sum E[K_b(x - x_i)]$$

$$= \frac{1}{n} \sum \int K_b(x - x_i) f_X(x_i)\, dx_i = \frac{1}{n} \sum K_b * f_X$$

b越小,和真实分布越相似,容易overfit

期望是Kb和f(x)的卷积

$$\text{Bias} = K_b * f_X(x) - f_X(x)$$

b越大bias越大,var越小

$$\text{VAR} = \frac{1}{n} \text{Var}[K_b(x - x_i)]$$

$$\text{MSE} = \text{Bias}^2 + \text{VaR}$$

$$\text{IMSE}(\hat{f}_b) = \int \text{MSE}(\hat{f}_b(x))\, dx$$

Remark. Implications of result on previous slide is:

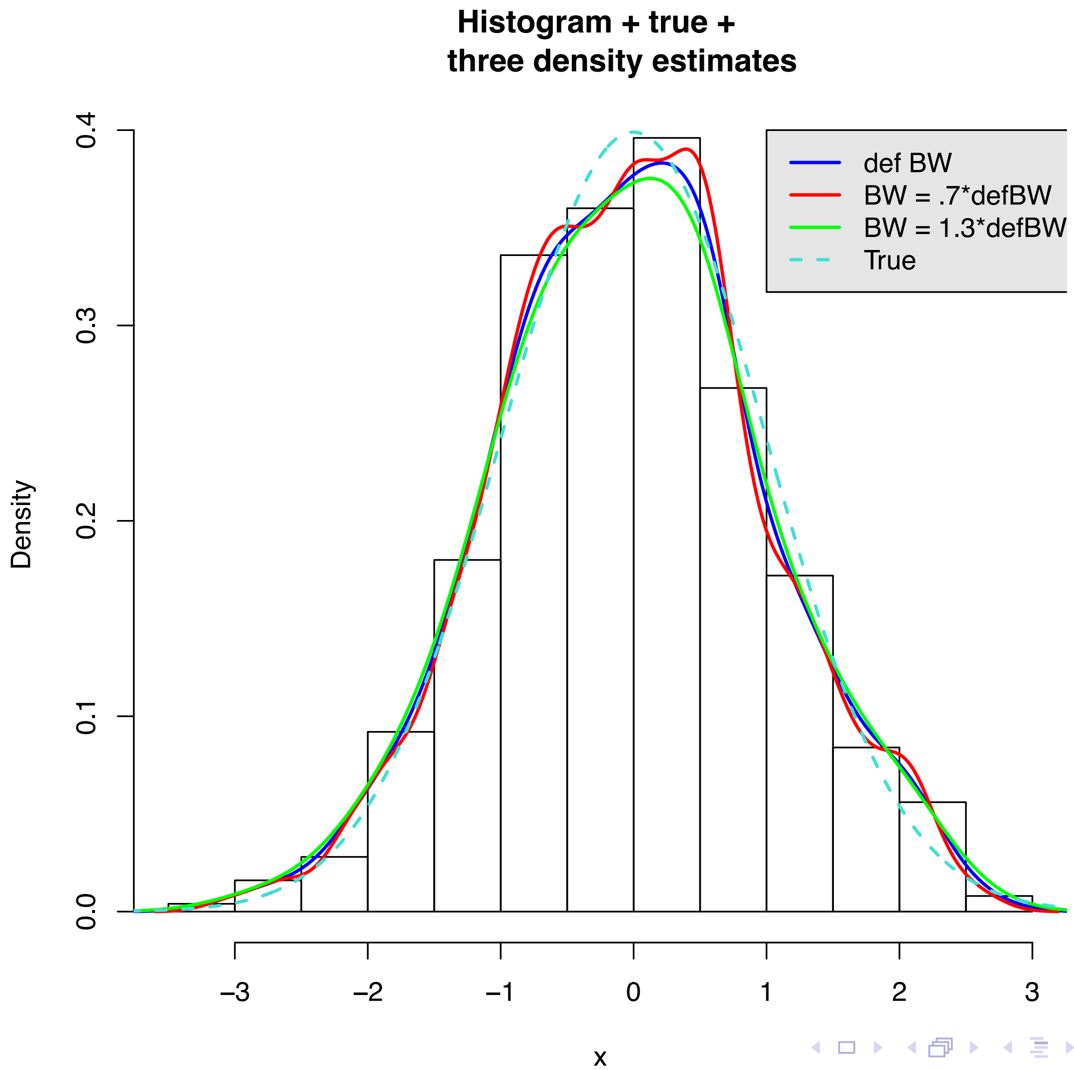
> The density estimates are estimated smoothed version of original distribution, smoothing depends on bandwidth

Example R-commands for plotting density estimate :

```
> plot(density(x))
> show(density(xnorm)$bw)
[1] 0.35056
> plot(density(x,bw=.4,kernel=c("gaussian"))
```

- `x` is the sample vector
- bandwidth `bw` is "effective" standard deviation of kernel K_b
- default kernel is Gaussian
- default bandwidth chosen according to Silverman rule:
  - 0.9 times the min of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power

Density Estimation - Simulated Data

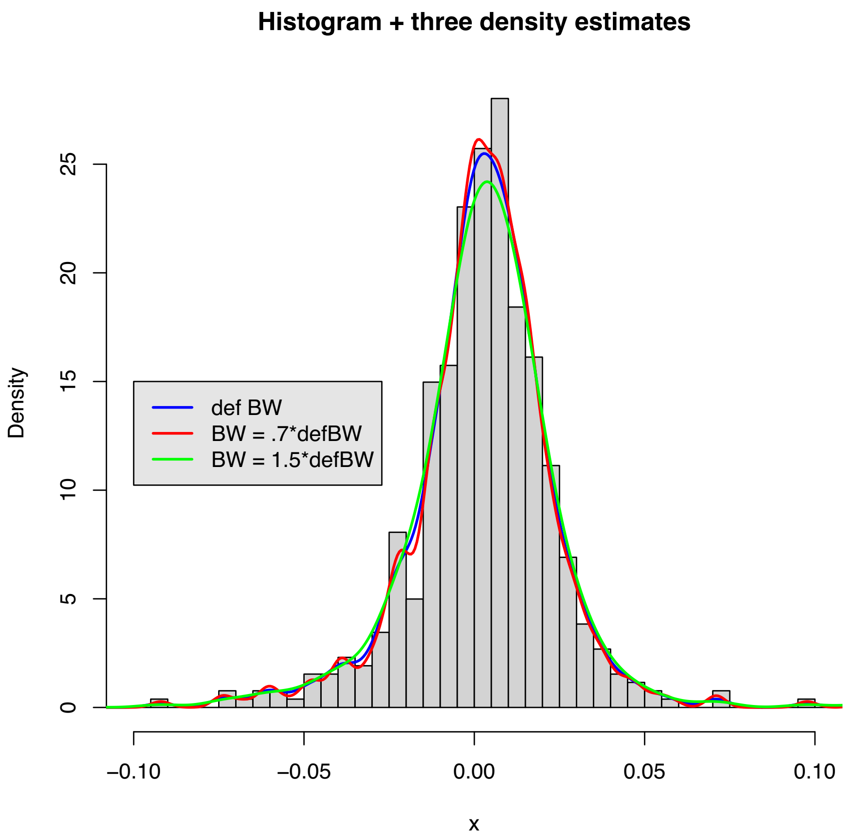


R-code

```
> xnorm <- rnorm(500,0,1)
> hist(xnorm,xlab='x',breaks = 20,freq=FALSE,main='Histogram + true +
+      three density estimates')
> # density estimate with default bw
> dens_dfbw = density(xnorm,kernel=c("gaussian"))
> # density estimates with adjusted bw
> dens_dfbw_dec = density(xnorm,kernel=c("gaussian"),adj=.7)
> dens_dfbw_inc = density(xnorm,kernel=c("gaussian"),adj=1.3)
> lines(dens_df,lty=1,lwd=2,col='blue')
> lines(dens_dfbw_dec,lty=1,lwd=2,col='red')
> lines(dens_dfbw_inc,lty=1,lwd=2,col='green')
> lines(seq(-4,4,by=.01),dnorm(seq(-4,4,by=.01),0,1),lty=2,lwd=2,
+       col='turquoise')
> legend(1,.4, c("def BW","BW = .7*defBW","BW = 1.3*defBW","True"),
+ lty=c(1,1,1,2),lwd=2, col=c("blue","red","green","turquoise"), bg="gray90")
> # show bandwidths
> show(c(dens_dfbw$bw,dens_dfbw_dec$bw,dens_dfbw_inc$bw))
[1] 0.2560331 0.1792232 0.3328431
```

Density Estimation - SP500 Weekly Log Returns

- Density estimation on the Weekly log returns for SP500 (Jan 2011 to Dec 2020)



R-code

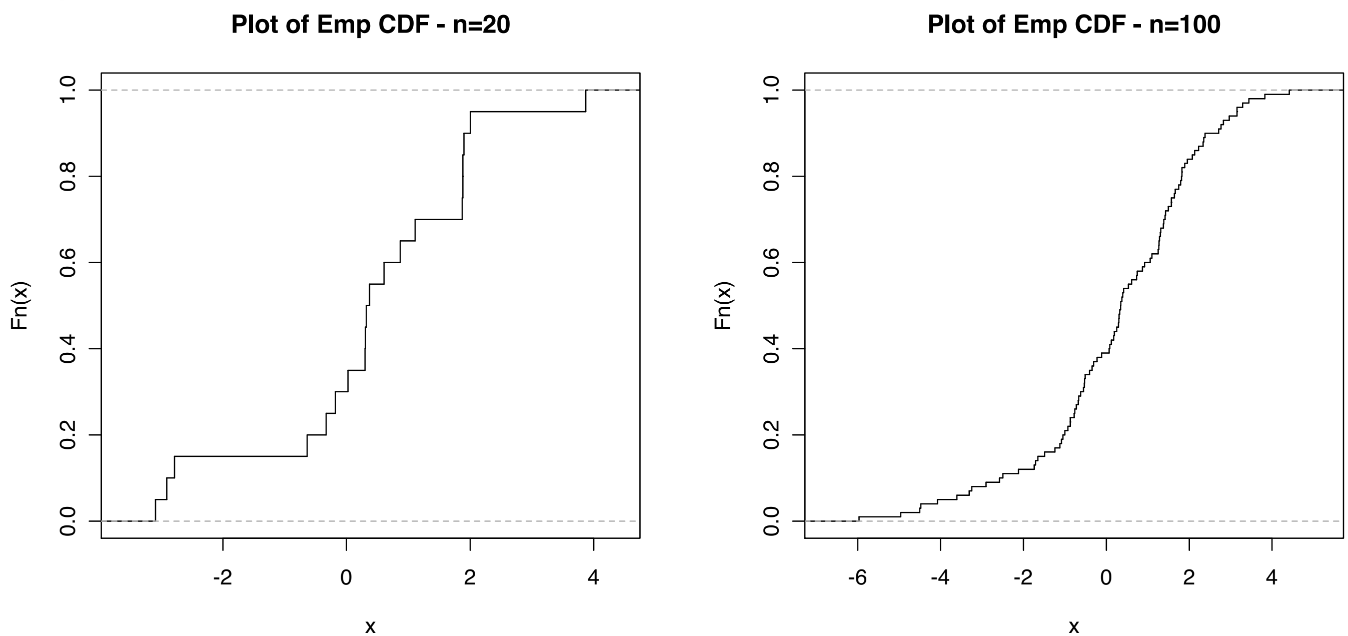
```
> X = read.csv("..\\Data\\SP500_Jan2011_Dec2020.csv",header=TRUE)
> SP500wk <- X$Adj.Close
> SP500wk_lret <- diff(log(SP500wk)) # generating log returns (weekly)
> windows()
> hist(SP500wk_lret,xlab='x',breaks = 40,xlim=c(-.10,.10),freq=FALSE,
+ main='Histogram + three density estimates')
> # density estimate with default bw
> SP500dens_dfbw = density(SP500wk_lret,kernel=c("gaussian"))
> # density estimates with adjusted bw
> SP500dens_dfbw_dec = density(SP500wk_lret,kernel=c("gaussian"),adj=.7)
> SP500dens_dfbw_inc = density(SP500wk_lret,kernel=c("gaussian"),adj=1.5)
> lines(SP500dens_dfbw,lty=1,lwd=2,col='blue')
> lines(SP500dens_dfbw_dec,lty=1,lwd=2,col='red')
> lines(SP500dens_dfbw_inc,lty=1,lwd=2,col='green')
> legend(-.1,15, c("def BW","BW = .7*defBW","BW = 1.5*defBW"), lty=c(1,1,1,2),
+ lwd=2, col=c("blue","red","green","turquoise"), bg="gray90")
> # show bandwidths
> show(c(SP500dens_dfbw$bw,SP500dens_dfbw_dec$bw,SP500dens_dfbw_inc$bw))
[1] 0.003971674 0.002780172 0.005957511
```

Empirical Distribution Function

Definition. With data $x_1, x_2, \ldots, x_n$, the empirical distribution function is defined as

$$\hat{F}_n(x) = \frac{1}{n} \# \{i : x_i \leq x\}$$ 离散CDF $= \frac{1}{n} \mathrm{Binom}(n, F(x))$

Example. Simulated data from $\mathcal{N}(0, 2^2)$ with two different sample sizes.



R-Code

```
x <- rnorm(20,0,2)
plot(ecdf(x), verticals=TRUE, do.p=FALSE, main='Plot of Emp CDF - n=20')

windows()
x <- rnorm(100,0,2)
plot(ecdf(x), verticals=TRUE, do.p=FALSE, main='Plot of Emp CDF - n=100')
```

Quantiles/Sample Quantiles

Recall. For distribution F, let π_q denote the q-quantile.

Definition. For sample of $x_1, x_2, \ldots, x_n$, the sample q-quantile is (simply) the q-quantile of the empirical CDF, i.e., the value $\hat{\pi}_q$ such that

$$\hat{\pi}_q = \hat{F}_n^{-1}(q) = \inf\{x : \hat{F}(x) \geq q\}$$

The **sample median** is $\hat{\pi}_{0.5}$. – interpretation 中位数

Remark. For random sample, $\hat{\pi}_q$ is an estimate of π_q. consistent estimate biased

Remark. If $x_1, x_2, \ldots, x_n$ is sample, the order statistics are the rearrangement of the values from smallest to largest, i.e.,

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

Then $x_{(i)}$ is the π_q empirical quantile if $\frac{i-1}{n} \leq q \leq \frac{i}{n}$

Background: QQ-Plots and Tailplots

Typical Problem. Suppose $x_1, x_2, \ldots, x_n$ is a sample from some process – interested in what is the appropriate parametric distribution/pdf. Use for estimating

- Parameters (e.g., mean and variance)
- Quantiles

Remark. Often the main focus is on the tail distribution – what is the probability of a loss exceeding some value? Relates to **Value-at-Risk, VaR** .

Q-Q Plots

Background:

- Comparing two distributions: plotting quantiles of one distribution against the corresponding quantiles of another distribution
- Common application is plots of (empirical) quantiles $\hat{F}_n^{-1}(q)$ vs. the quantiles of the "estimated" cdf $F^{-1}(q)$ at n equally spaced quantile values of

$$q = \frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$$

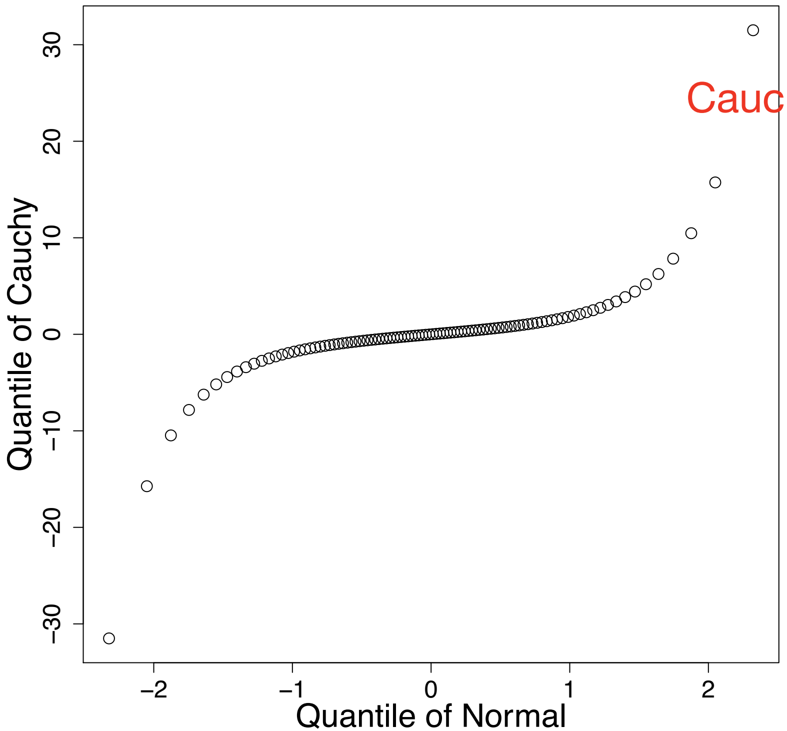
  - Requires an estimation step, i.e., estimating parameters
    - Utilize well-accepted estimation methodology – typically maximum-likelihood or some modification

Remark. Q-Q plots are equivalent to plotting the order statistics $x_{(k)}$ vs. $F^{-1}\left(\frac{k}{n+1}\right)$.

Normal vs Cauchy

这是两个分布比较的QQ plot

```
> q1 <- qnorm(seq(0, 1, length=100), mean=0, sd=1)
> q2 <- qcauchy(seq(0, 1, length=100),
    location=0, scale=1)
> plot(q1, q2, xlab="Quantile of Normal",
    ylab="Quantile of Cauchy")
```

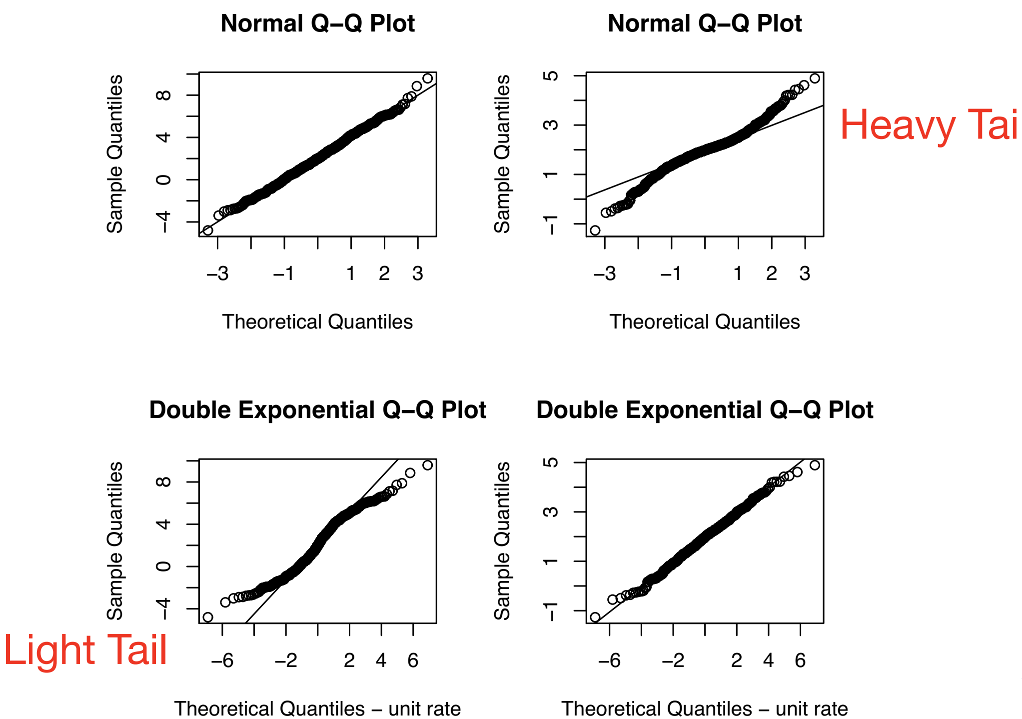


Cauchy has a heavy tail

QQ Plots - Simulated Data

Example. Simulated 1000 random deviates from $\mathcal{N}(2, 2^2)$ and a 1000 random deviates from DExp$(2, 2)$. Generated

- normal Q-Q plots for both
- double exponential Q-Q plots for both
- QQ plots in left column are for the normal data
- QQ plots in right column are for the double exponential data



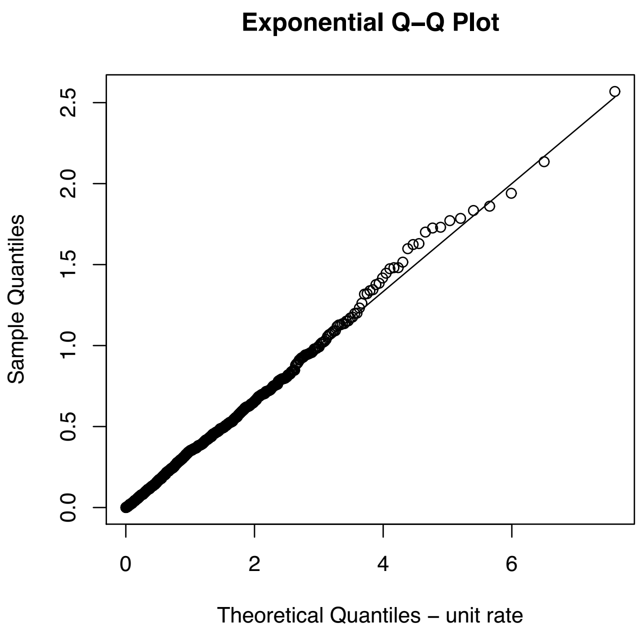
R-code for QQ Plots

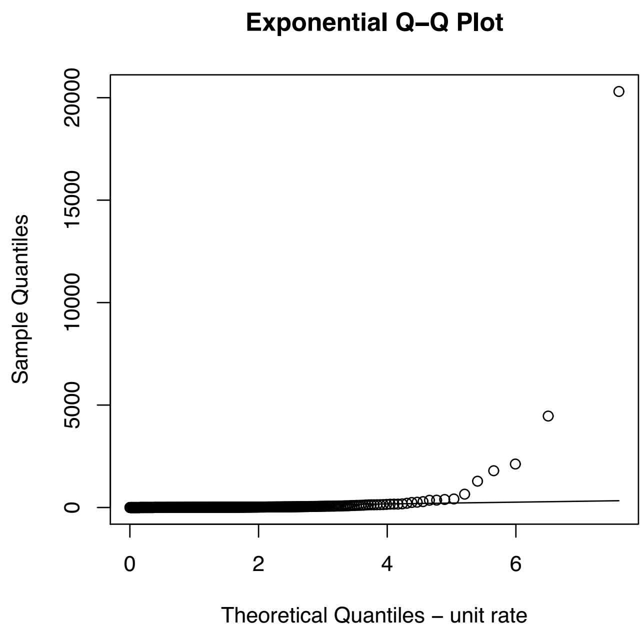
```
xnorm <- rnorm(1000,2,2)
xdexp <- rdexp(1000,2,2)

windows()
par(mfrow=c(2,2)) # setting up 2 x 2 arrangement of subplots
qqnorm(xnorm)
qqline(xnorm)
qqnorm(xdexp)
qqline(xdexp)
qqdexp(xnorm)
qqdexp(xdexp)
```

QQ Plots - Simulated Data II

Example. Simulated 1000 random deviates from Exp(3) and 1000 random deviates from GPD$(1, 0, 3)$.

- Generated Q-Q plots for exponential distibution – applied to both data sets
  - On the left is plot for exponential "data"
  - On the right is plot for generalized pareto "data"





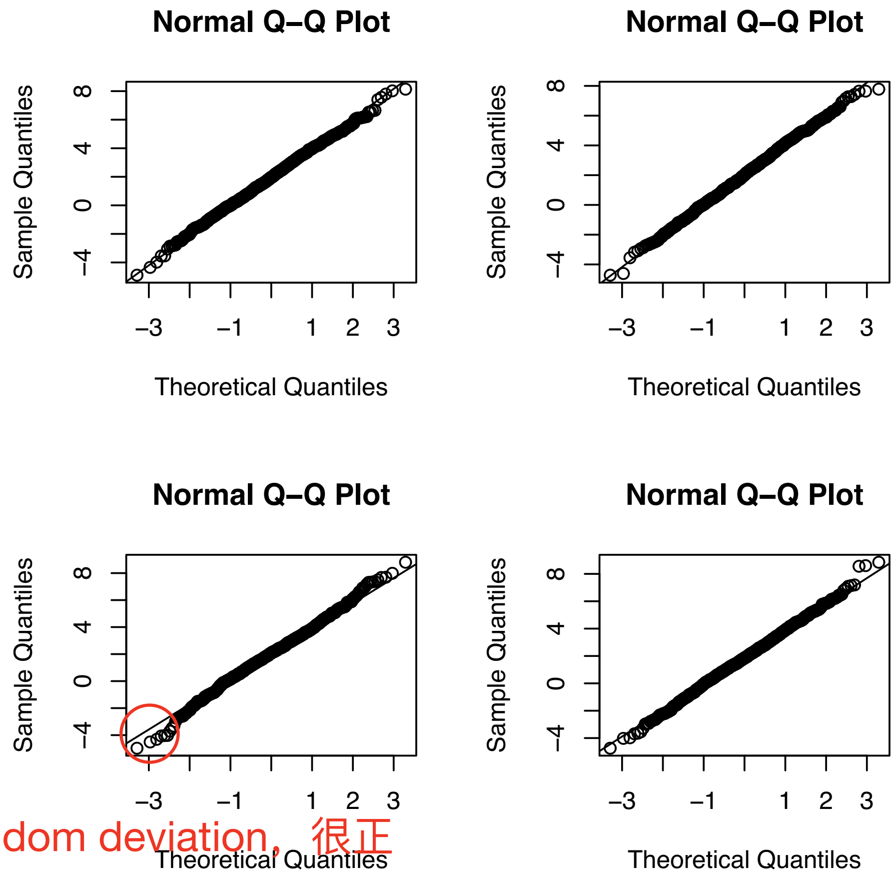
Interpretation.

R-code for Exp/Pareto QQ-Plots

```
xexp <- rexp(1000,3)
xgpd <- rgpd(1000,1,0,3)
windows()
qqexp(xexp)
windows()
qqexp(xgpd)
```

QQ Plots - Simulated Data III

Example. Simulated 1000 random deviates from $\mathcal{N}(2,2)$ – did this 4 different times. Note the randomness in the plots.



Random deviation，很正常,但需要进行test

R-Code for Normal QQ-plots

```
windows()
par(mfrow=c(2,2)) # setting up 2 x 2 arrangement of subplots
for(i in 1:4) {
x<- rnorm(1000,2,2)
qqnorm(x)
qqline(x)
}
```

PCS Data

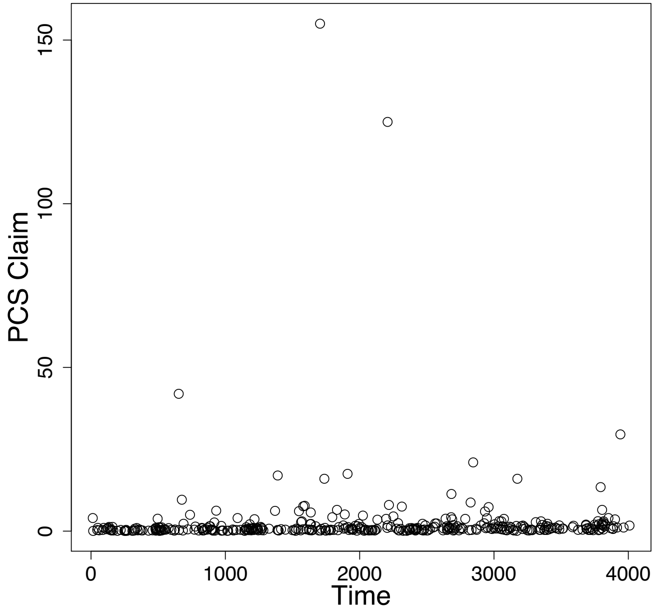
Background. Product Claim Services (PCS) is a division of ISO

- ISO basically is a global company developing tools/data for analyzing/quantifying risk in a wide variety of applications
- PCS gathers data for total insurance claims on catastrophes
  - Currently defined to be claims of \$25 million or more
  - Data has claims down to \$7 million
- Options and futures contracts on the PCS Index offer a possibility to securitize insurance catastrophe risk.

`http://www.iso.com/index.php?option=com_content&task=view&id=743`

PCS Data – Loading in R

```
## Load the data
> load("PCS.rda")
## Check out the data
> PCS
     Col1   Col2
1      13   4.00
2      16   0.07
3      46   0.35
4      60   0.25
...
> plot(PCS[,1], PCS[,2], xlab="Time", ylab="PCS Claim")
```

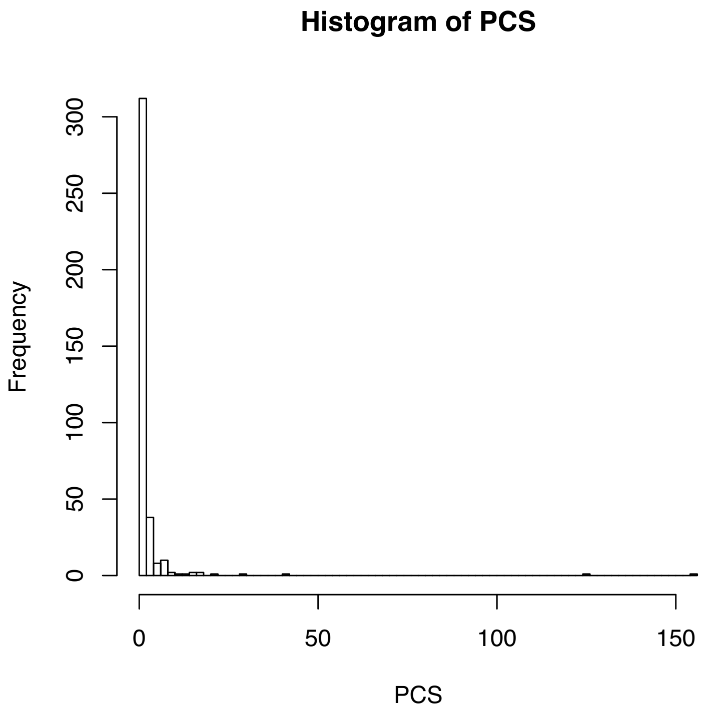


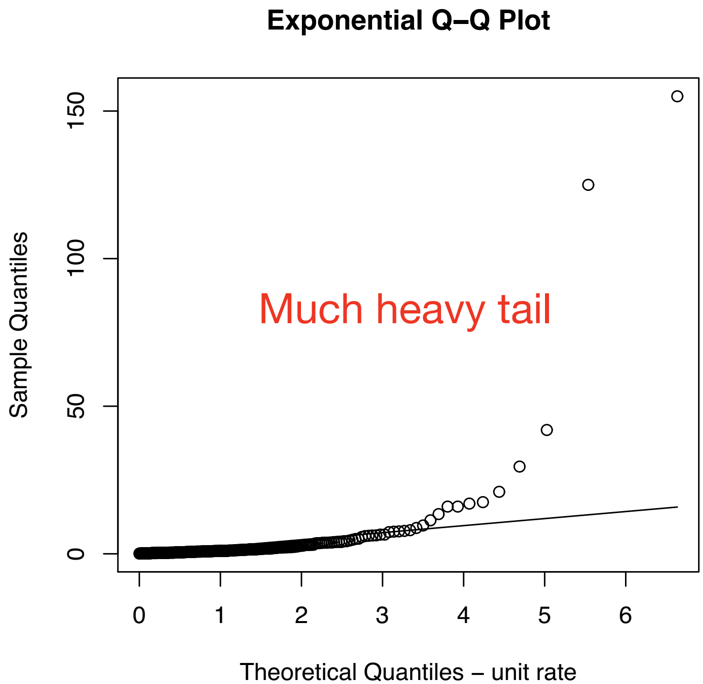
- First column is time stamp – corresponding to day
- Second column is the claim (in 100 million dollars)

Distributional Analysis of PCS Claims Data

Histogram and QQ Plot relative to Exponential

Interpretation



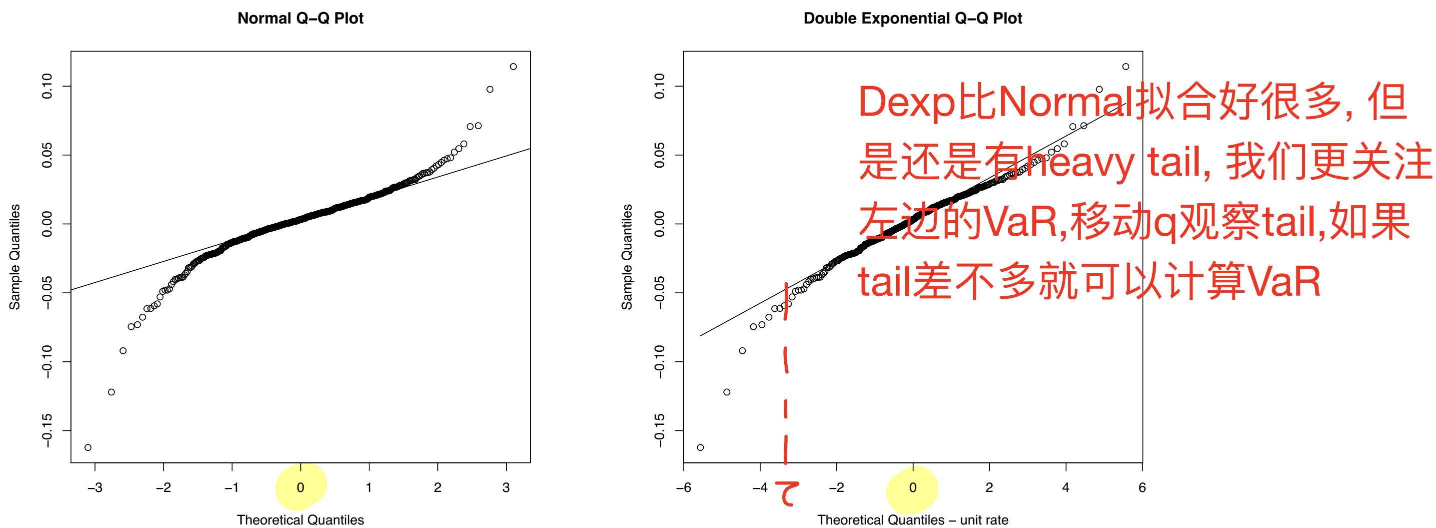


QQ Plots - SP500

Remark. Generated QQ plots of SP500 wkly log returns

- Normal and Double Exponential QQ Plots

Motivation/Interpretation




```
> qqnorm(SP500wk_lret)
> qqline(SP500wk_lret)
> qqdexp(SP500wk_lret)
> kurtosis(SP500wk_lret)
[1] 8.935352
attr(,"method")
[1] "excess"
```

只和standard normal distribution作QQ plot,而不是估计均值和方差之后再作图,寻找linearity

Tests of Normality

Best

- Shapiro-Wilk (Focused on QQ-Plot Analysis) 首选的Test,因为关注tail

不好
- Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises (comparison between theoretical cdf and empirical cdf) 一般 一般 test goodness of fit不好用
- Jarque-Bera (Weighted sum of Skewness and Kurtosis)

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

where

- S is empirical skewness parameter (of est residuals)
- K is empirical kurtosis parameter (of est residuals)
- Under the null distribution (residuals are normally distributed), the approximate distribution of JB is approximately chi-square with 2 degrees of freedom

Reject Normality (of errors) if JB is large

R-Functions for Tests of Normality

- Shapiro-Wilk test - `shapiro.test(x)`
- Jarque-Bera

R语言默认只能test normal

```
rjb.test {lawstat} R Documentation
Test of Normailty - Robust Jarque Bera Test

Description: This function performs robust & classical Jarque-Bera tests of
             normality.

Usage: rjb.test(x, option = c("RJB", "JB"),
         crit.values = c("chisq.approximation", "empirical"), N = 0)

Arguments:
x  a numeric vector of data values.
option  The choice of the test must be "RJB" (default) or "JB".
crit.values  character string specifying how critical values are obtained,
             i.e. approximated by chisq-distribution (def) or empirically.
N  number of Monte Carlo simulations for empirical critical values
```

Tests of Normality on SP500 Weekly Log Returns

```
> shapiro.test(SP500wk_lret)

Shapiro-Wilk normality test

data:  SP500wk_lret
W = 0.89558, p-value < 2.2e-16

> rjb.test(SP500wk_lret)

Robust Jarque Bera Test

data:  SP500wk_lret
X-squared = 3989.9, df = 2, p-value < 2.2e-16
```

not norma

Tail Analysis of Extreme Distributions

Remarks.

- QQ Plots shown so far are showing the fit relative to the whole distribution
- Have shown example (SP500 log returns) where we analyzed the postive and negative returns separately
- Interest in a more detailed analysis of the tail distribution – trying to answer questions of

Question 1: What is the appropriate model for the tail distribution

- To define "tail" utilize a threshold τ, i.e., the tail $1 - F(x)$ for $x \geq \tau$

Question 2: Is the tail distribution (model) consistent for a range of threshold values τ?

Tail Analysis of Extreme Distributions

Remark. To help answer the questions, there are a number of techniques that are useful – two we cover are

- Estimation of distribution parameters based on data values larger than specified threshold
    - Tailplot comparison with empirical data
- Plot of the estimated shape parameter as a function of threshold

  xi不会改变
    - Would like it to be consistent
    - Provides some guidance on appropriate thresholds to use in estimation

Remark. There are a number of other "extreme" distributions

- We only cover generalized pareto
- Techniques presented here can be applied to these other distributions

Pareto Distribution

Pareto is the best choice for everything

- Density

$$f_{a,\mu}(x) = \frac{a\mu^a}{x^{1+a}}, \quad x > \mu$$

where a is called the shape parameter, or shape index of the tail. The density of the distribution decays polynomially. (Due to Swiss economist Vilfredo Pareto)

- CDF

$$F_{a,\mu}(x) = \begin{cases} 0 & \text{if } x < \mu \\ 1 - \left(\frac{\mu}{x}\right)^a & \text{if } x \geq \mu \end{cases}$$

- Mean – $E(X) = \frac{a\mu}{a-1}$, $a > 1$.
- Variance – $\text{Var}(X) = \frac{a\mu^2}{(a-1)^2(a-2)}$, $a > 2$.

Generalized Pareto Distribution (GPD)

- Density

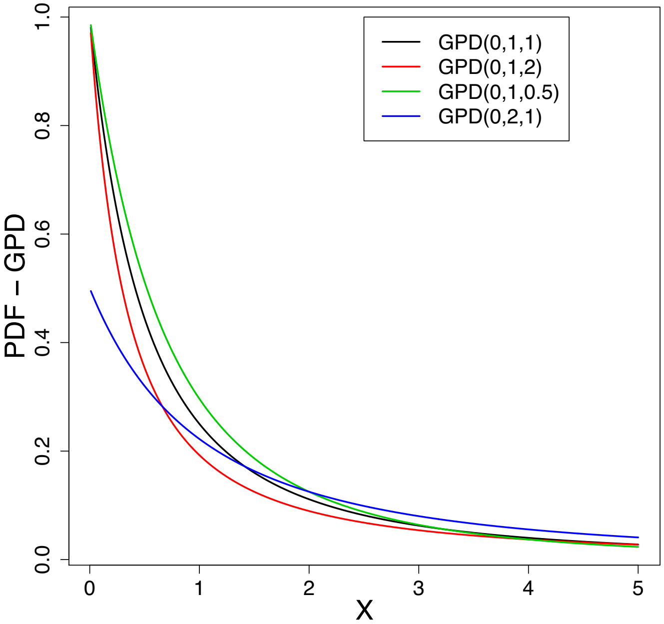
$$f_{\mu,\sigma,\xi}(x) = \frac{1}{\sigma}\frac{1}{(1+\xi(x-\mu)/\sigma)^{1+1/\xi}}, \quad x > \mu$$

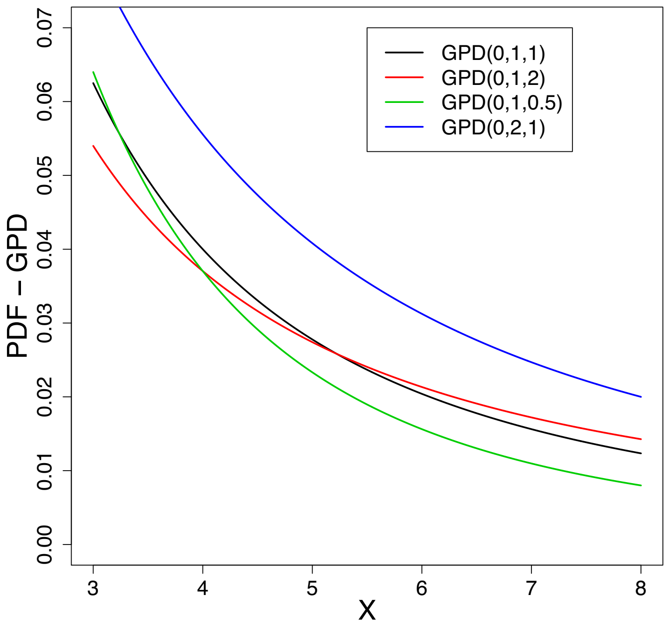
- CDF

$$F_{\mu,\sigma,\xi} = \begin{cases} 0 & \text{if } x < \mu \\ 1 - \frac{1}{(1+\xi(x-\mu)/\sigma)^{1/\xi}} & \text{if } x \geq \mu \end{cases}$$

- Pareto and GPD are equal when $\xi = 1/a$ and $\sigma = \mu/a$.
- Exponential distribution: $\xi = 0$ and $\mu = 0$.

```
> x <- seq(0.01, 5, length=1000)
> plot(x, dgpd(x, m=0, lambda=1, xi=1),
  xlab="X", ylab="PDF - GPD", type="l",
  col=1, lty=1)
> lines(x, dgpd(x, m=0, lambda=1, xi=2),
  col=2, lty=1)
> lines(x, dgpd(x, m=0, lambda=1, xi=0.5),
  col=3, lty=1)
> lines(x, dgpd(x, m=0, lambda=2, xi=1),
  col=4, lty=1)
> legend(2.5, 1, legend=c("GPD(0,1,1)", "GPD(0,1,2)",
  "GPD(0,1,0.5)", "GPD(0,2,1)"), lty=1, col=c(1,2,3,4))
```


GPD(0,1,1)
GPD(0,1,2)
GPD(0,1,0.5)
GPD(0,2,1)
PDF – GPD
X


GPD(0,1,1)
GPD(0,1,2)
GPD(0,1,0.5)
GPD(0,2,1)
PDF – GPD
X

Pickands-Balkema-de Haan Theorem I

Adapted from Theorem 3.4.5 in **Embrechts, Kluppelberg and Mikosch**.

Theorem (Pickands-Balkema-de Haan)

Under mild conditions on the distribution of rv X, we have

X>0

$$\frac{\mathbb{P}(X > u + x\,a(u))}{\mathbb{P}(X > u)} \longrightarrow H_\xi(x), \quad \textit{as } u \uparrow \infty.$$

for some $a(u) > 0$, where

$$H_\xi(x) = \Big(1 + \xi x\Big)^{-1/\xi}, \quad \textit{for } 1 + \xi x > 0.$$

Interpretation: 对于u,可以找到a(u)转换把大部分分布转换为Pareto分布 $\sigma=1$

Pickands-Balkema-de Haan Theorem II

a必定存在,但不需要知道具体值,因为它确保了X是Pareto分布

- Suppose $x > 0$ and $\xi > 0$. Then, as $u \uparrow \infty$,

$$\frac{\mathbb{P}(X > u + x a(u))}{\mathbb{P}(X > u)} \to \frac{1}{(1+\xi x)^{-1/\xi}}.$$

- Since $a(u) > 0, x > 0$, the left-hand-side equals

$$P[X > u + x a(u) \mid X > u] = P\left[\frac{\cancel{X} - u}{a(u)} > x \,\middle|\, \cancel{X} > u\right] \longrightarrow \frac{1}{(1+\xi x)^{-1/\xi}}$$

由于Pareto的linearity,X也是Pareto

- This means that, for large u, the conditional distribution of the excess loss $X - \mu$, given that the loss is greater than u is approximately $\frac{1}{\xi}$