Lecture 3: Univariate Descriptive Statistics/EDA

Brian Thelen
443 West Hall
bjthelen@umich.edu

Statistics 509 - Winter 2016
Ref: Ruppert: Chapter 4.1-4.5

Overview of Lecture.

- Descriptive statistics
  - Summary quantitative statistics
  - Graphical summaries
    - histograms
    - density estimation
    - boxplots
- 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 2

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
- 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
  - Expected "closeness" of hte 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 and kurtosis is
- 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 and , 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
```

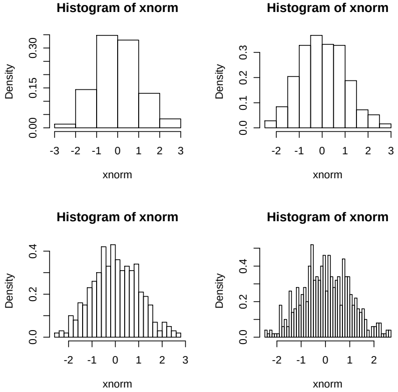
```
> X = read.csv("Data\\SP500_wkly_Jan1_60_Jul23_07.csv",header=TRUE)
> SP500wk <- rev(X\$Close)
> SP500wk_lreturn <- diff(log(SP500wk)) # log returns (weekly)
> skewness(SP500wk_lreturn)
[1] -0.3662413
> kurtosis(SP500wk_lreturn)
[1] 3.387008
```

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)



Question. Which one is preferred?

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.

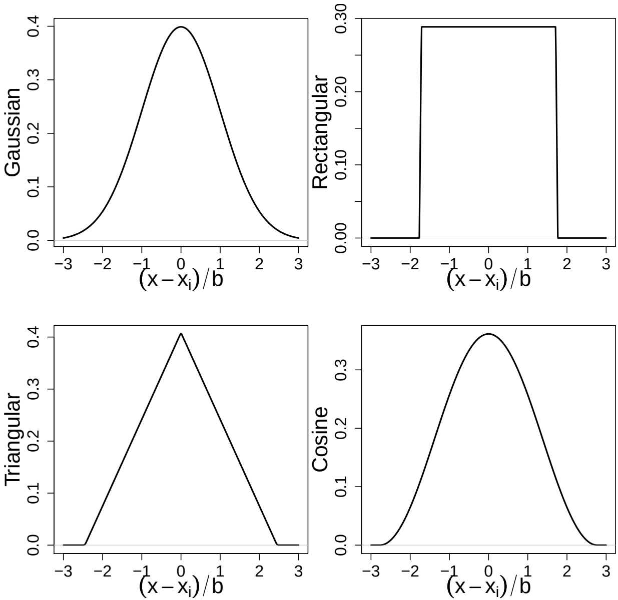
Definition. For sample data $x_1, x_2, \ldots, x_n$, a kernel-based density estimate is defined as

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

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
  - Possible shapes for K are "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" or "optcosine"
- 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 defintions of bandwidth parameter – larger BW corresponds to more "smoothing" (i.e., bias in estimation) and less "noise" (i.e., variance in estimation).

Remark. Effect of bandwidth is

- When bw parameter b gets small,

- When bw parameter b gets large,

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

Answer.

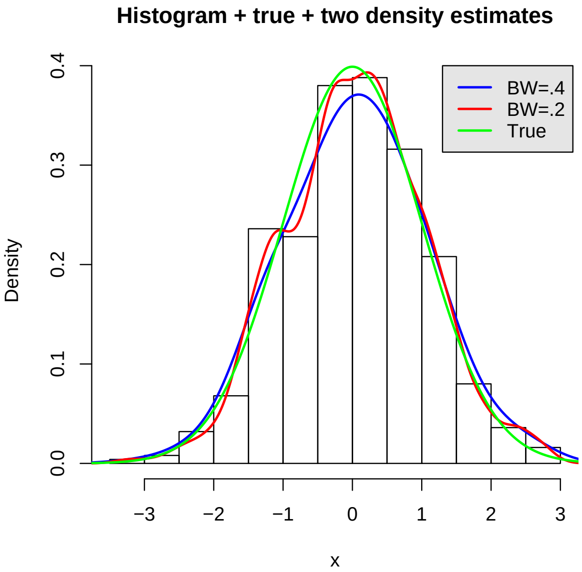
Remark. Implications of result on previous slide is:

R-command for plotting density estimate :

```
plot(density(x,bw=.4,kernel=c("gaussian"))
```

- `x` is the sample vector
- bandwidth `bw` is "effective" standard deviation of kernel K_b

Density Estimation - Simulated Data



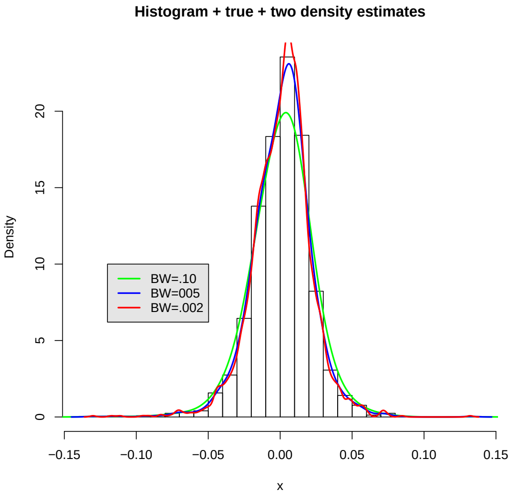
R-code

```
xnorm <- rnorm(500,0,1)
windows()
hist(xnorm,xlab='x',breaks = 20,freq=FALSE,main='Histogram + true +
      two density estimates')
lines(density(xnorm,bw=.4,kernel=c("gaussian")),lty=1,lwd=2,col='blue')
lines(density(xnorm,bw=.2,kernel=c("gaussian")),lty=1,lwd=2,col='red')
lines(seq(-4,4,by=.01),dnorm(seq(-4,4,by=.01),0,1),lty=1,lwd=2,
      col='green')
legend(1.3,.4, c("BW=.4","BW=.2","True"), lty=1,lwd=2,
      col=c("blue","red","green"), bg="gray90")
```

- Density estimates with Gaussian kernel (this is default)

Density Estimation - SP500 Data

- Density estimation on the log(weekly return) for SP500



R-code

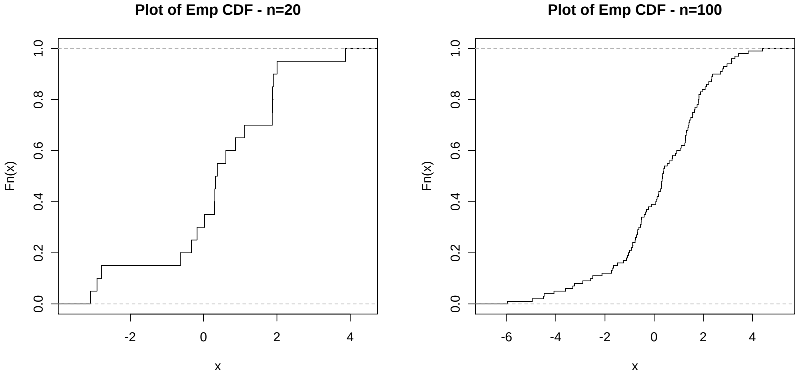
```
windows()
hist(SP500wk_lreturn,xlab='x',breaks = 20,freq=FALSE,main=
      'Histogram + true + two density estimates')
lines(density(SP500wk_lreturn,bw=.010,kernel=c("gaussian")),
      lty=1,lwd=2,col='green')
lines(density(SP500wk_lreturn,bw=.005,kernel=c("gaussian")),
      lty=1,lwd=2,col='blue')
lines(density(SP500wk_lreturn,bw=.002,kernel=c("gaussian")),
      lty=1,lwd=2,col='red')
legend(-.12,10, c("BW=.10","BW=005","BW=.002"), lty=1,lwd=2,
      col=c("green","blue","red"), bg="gray90")
```

Empirical Distribution Function

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

$$\hat{F}_n(x) = \frac{1}{n} \# \{i : \; x_i \leq x\}$$

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

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

The **sample median** is $\hat{\pi}_{.50}$ – interpretation

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

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.,

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

Example.

Sample
$x_1 = 30,\ x_2 = 60,\ x_3 = 10,\ x_4 = 100,\ x_5 = 30$
Order Statistics
$x_{(1)} =$, $x_{(2)} =$, $x_{(3)} =$, $x_{(4)} =$, $x_{(5)} =$

Question. What is the relationship between sample quantiles and the order statistics?

Answer.

Boxplots

Definition The inter-quartile range (IQR) is the difference between the first and third quartiles, $\hat{\pi}_{.25}, \hat{\pi}_{.75}$ i.e.,

$$\text{IQR} = \hat{\pi}_{.75} - \hat{\pi}_{.25}$$

Note that the IQR is the range of the middle 50% of the data.

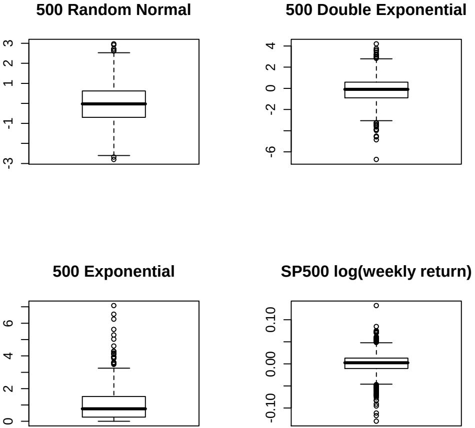
- An sample value is labeled an **outlier** if it lies (at least) $1.5 \cdot \text{IQR}$ below $\hat{\pi}_{.25}$ or above $\hat{\pi}_{.75}$.

Details for Boxplots

(1) Box with sides going from $\hat{\pi}_{.25}$ to $\hat{\pi}_{.75}$.
(2) Line in box at the median.
(3) Draw lines out furthest observations within $1.5 \cdot \text{IQR}$ of edges of box.
(4) Put "o" at the **outlier** values that are more than $1.5 \cdot \text{IQR}$ from the edges of the box.

Example Box Plots

- Box plot of 500 random deviates from $\mathcal{N}(0, 1)$
- Box plot of 500 random deviates from $\text{DExp}(0, 1)$
- Box plot of 500 random deviates from $\text{Exp}(1)$
- Box plot of SP500 log(weekly returns)



R-code for Boxplots

```
# Boxplot Examples
xnorm <- rnorm(500,0,1)
xdexp <- rdexp(500,0,1)
xexp <- rexp(500,1)

par(mfrow=c(2,2)) # setting up for a 2 x 2 arrangement of subplots

boxplot(xnorm)
title('500 Normal')

boxplot(xdexp)
title('500 Double Exponential')

boxplot(xexp)
title('500 Exponential')

boxplot(SP500wk_lreturn)
title('SP500 log(weekly return)')
```

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

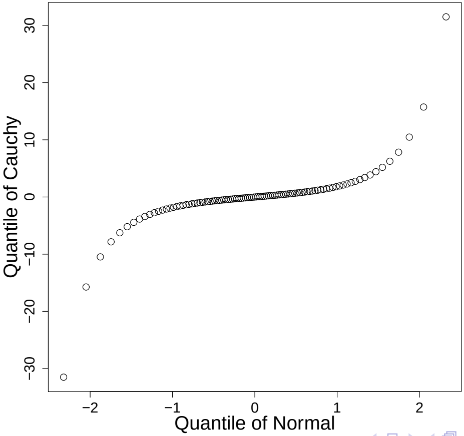
$$q = \frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$$

  - 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

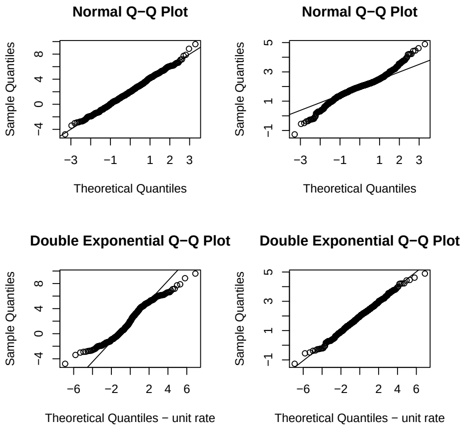
```
> q1 <- qnorm(seq(0, 1, length=100), mean=0, sd=1)
> q2 <- qcauchy(seq(0, 1, length=100),
    location=0, scale=1)
> plot(q1, q2, xlab="Quantile of Normal",
    ylab="Quantile of Cauchy")
```



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



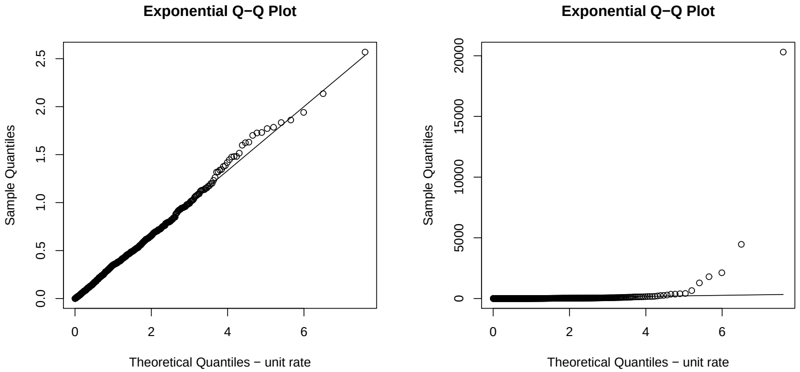
R-code for QQ Plots

```
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 $\text{Exp}(3)$ and 1000 random deviates from $\text{GPD}(1, 0, 3)$.

- 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"



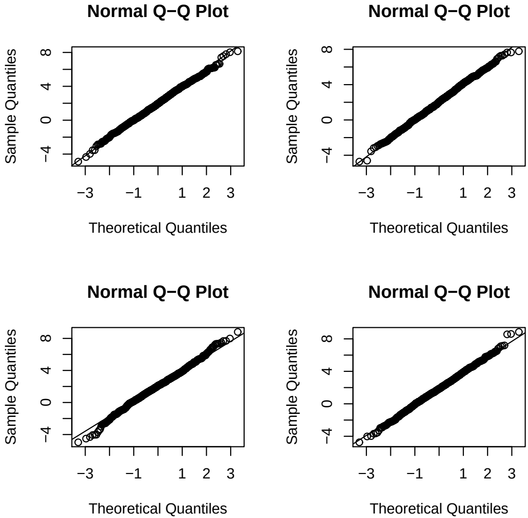
Interpretation.

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.



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

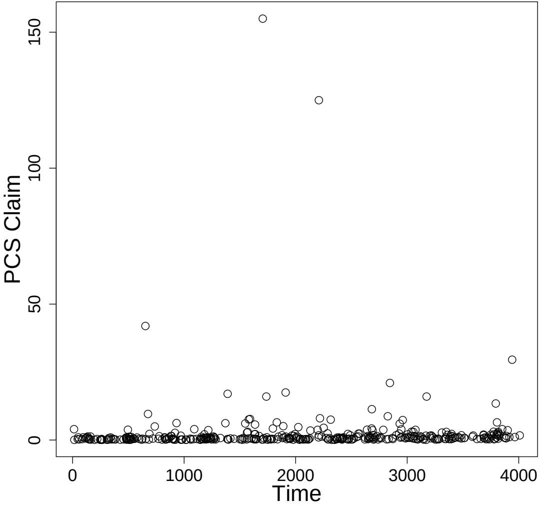
- ISO basically is a global company developing tools/data for analyzing/quantifying risk in a wide variety of applications
- PCS gathers data for total insurance claims on catastrophes
  - Currently defined to be claims of $25 million or more
  - Data has claims down to $7 million
- Options and futures 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")
```

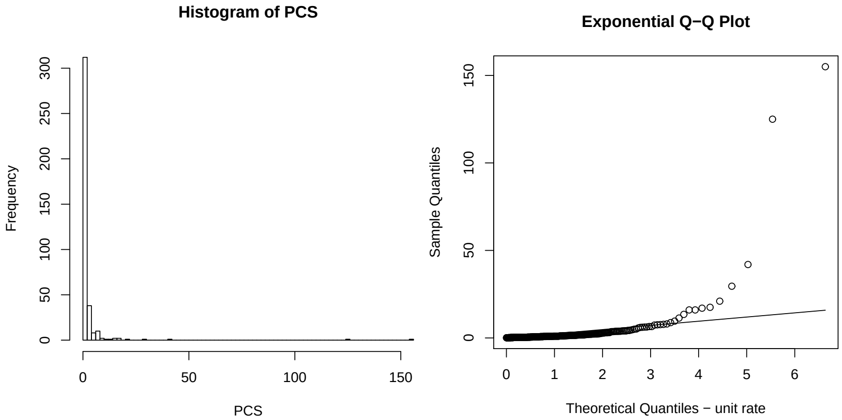
- First column is time stamp – corresponding to day
- Second column is the claim (in 100 million dollars)


PCS Claim
0
50
100
150
0
1000
2000
3000
4000
Time

Distributional Analysis of PCS Claims Data

Histogram and QQ Plot relative to Exponential

Interpretation

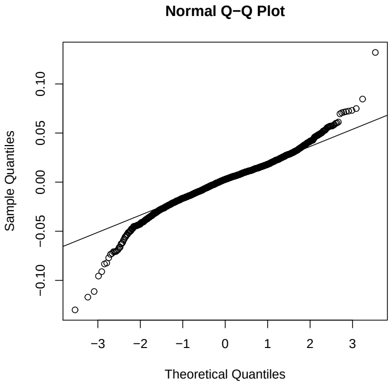


QQ Plots - SP500

Remark. Generated QQ plots of log(SP500 wkly returns)

- Normal QQ plot (using function `myqqnorm`)

Motivation/Interpretation



Tests of Normality

- Shapiro-Wilk (Focused on QQ-Plot Analysis)
- Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises (comparison between theoretical cdf and empirical cdf)
- 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

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

Description: This function performs robust & classical Jarque-Bera test

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 obtain
             i.e. approximated by chisq-distribution (def) or empirical
N  number of Monte Carlo simulations for empirical critical values
```

Tests of Normality on SP500 Weekly Log Returns

```
> X = read.csv("Data\\SP500_wkly_Jan1_60_Jul23_07.csv",header=TRUE)
> SP500wk <- rev(X$Close)
> SP500wk_lreturn <- diff(log(SP500wk)) # generating log returns (weekl
> shapiro.test(SP500wk_lreturn)

        Shapiro-Wilk normality test

data:  SP500wk_lreturn
W = 0.9678, p-value < 2.2e-16

> library(lawstat)
> rjb.test(SP500wk_lreturn)

        Robust Jarque Bera Test

data:  SP500wk_lreturn
X-squared = 1327.092, df = 2, p-value < 2.2e-16
```

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
  - 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

- 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

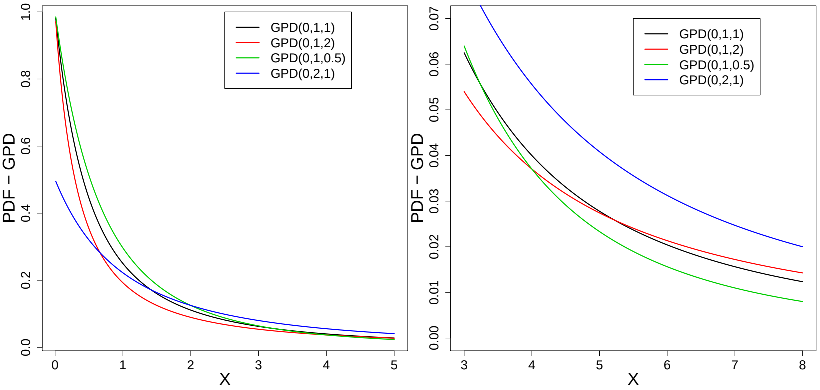
$$f_{\mu,\sigma,\xi}(x) = \frac{1}{\sigma}\frac{1}{(1+\xi(x-\mu)/\sigma)^{1+1/\xi}}, \quad x > \mu$$

- 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
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
GPD(0,1,1)
GPD(0,1,2)
GPD(0,1,0.5)
GPD(0,2,1)
PDF – GPD
X
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
3
4
5
6
7
8

Estimating the Shape Index

- Histograms and kernel density estimators can be good estimators in the center of a distribution where most of the data is to be found, but they are rather poor estimators of the tails.
- Peak over threshold (POT) methods
  - Pareto: Linear regression
  - GPD: Maximum likelihood estimation
- Issue of selecting the threshold.

Likelihood Function

- Probability models usually depend on unknown parameters θ (here θ can be a vector) – then the joint PDF of iid sample $x_1, \ldots, x_n$ can be written as

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

- We can view $L(\theta) = f(x_1, \ldots, x_n; \theta)$ as a function of θ with $x_1, \ldots, x_n$ fixed at the observed data, and we call it the likelihood function. It tells us the likelihood of the sample that was actually observed.

Maximum Likelihood Estimation

- Definition: The maximum-likelihood estimates (MLE) are the parameter values that maximize the likelihood function, or equivalently the values that maximize the log-likelihood function.
- The log-likelihood function is the (natural) logarithm of the above, i.e.,

$$\ell(\theta) = \log\left[L(\theta)\right]$$

Steps for Finding MLE

$$\max_{\theta} L(\theta) \quad \text{or} \quad \max_{\theta} \ell(\theta)$$

1. Derive the likelihood or log-likelihood function.
2. Take the derivative with respect to each of the parameters and set the derivatives equal to 0.
3. Solve for unknown parameters – these are the MLEs.

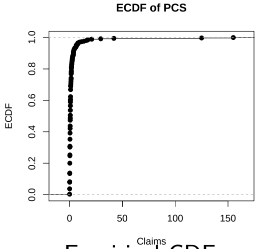
Peak over Threshold - Tail Fitting

Steps

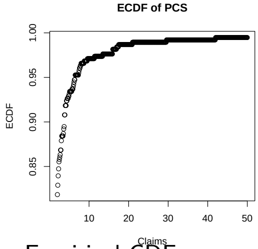
- Compute empirical cdf and look at the range of quantiles for the tail (.30 to .05 or less) – these are candidate thresholds
- Compute MLE of shape parameter (in the tail) using GPD model over range of threshold values and look at stability
  - Only estimating scale and shape, as the threshold is the location
- Pick threshold based on quantile considerations (not too far out in the tail), but also in the stable region relative to estimation, and compute MLEs of Generalized Pareto parameters
- Look at Goodness of Fit of the tail distribution relative to the estimated model – QQ plots of tails

PCS Index: POT Analysis

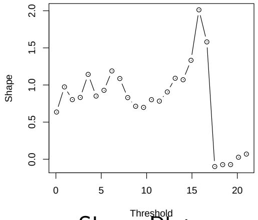
- Use R POT package for shape-plots/tail plots in POT tail distribution modeling - again using GPD model for tails



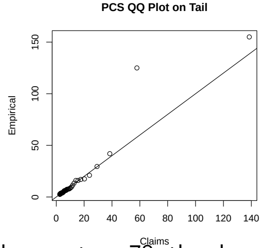
Empirical CDF



Empirical CDF - zoom



Shape Plots



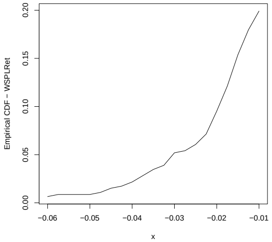
Shape est = .78, thresh = 2.5

R-commands for previous slide

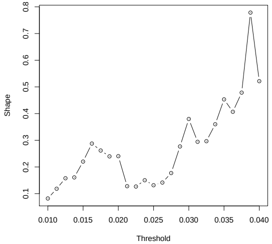
```
> library(POT)
> eecdf = ecdf(PCS[,2])
> plot(eecdf,main='ECDF of PCS',xlab='Claims',ylab='ECDF')
> uv = seq(from = 2,to = 50, by = .1)
> plot(uv,eecdf(uv),main='ECDF of PCS',xlab='Claims',ylab='ECDF')
> tcplot(PCS[,2],nt=25,conf=0)
> gpd_fit = fitgpd(PCS[,2],2.5)
> qq(gpd_fit, main='PCS QQ Plot on Tail', xlab='Claims', ylab='Empirical', ci =
> gpd_fit
Estimator: MLE
  Threshold Call: 2.5
    Number Above: 54
Proportion Above: 0.1421

Estimates
 scale   shape
2.7759  0.7851

Standard Errors
 scale   shape
0.6717  0.2269
Optimization Information
  Convergence: successful
  Function Evaluations: 26
  Gradient Evaluations: 13
```

SP500 Weekly Log Returns - POT

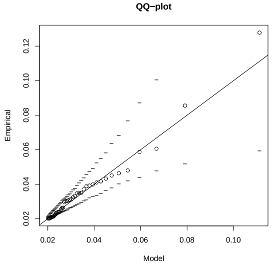
- Results for 1960-2007 data



Empirical CDF



Shape Plot



Tail Plot for $u = -.020$

R-Command - Using POT

```
> WSPLRet = SP500wk_lreturn
> library(POT)
> eecdf = ecdf(WSPLRet)
> uv = seq(from = -.06,to = -.01, by = .0025)

> plot(uv,eecdf(uv),type='l',xlab='x',ylab='Empirical CDF - WSPLRet')
>
> tcplot(-WSPLRet,c(.01,.04),nt=25,conf=0)
>
> gpd_fit = fitgpd(-WSPLRet,.02)
> qq(gpd_fit)
> scale = gpd_fit$fitted.values[1]
> xi = gpd_fit$fitted.values[2]
> xi
    shape
0.1225712
```

Computing VaR: Two Approaches

Problem: Have historical log-returns $X_1, \ldots, X_n$ (assuming stationary) with cdf F which is unknown and want to estimate VaR for a specified α (e.g., .01, .005) – this corresponds to estimating the α-quantile of the F, and taking the negative.

Two approaches: Nonparametric and Semiparametric

- Nonparametric approach – simply use sample quantile $\hat{\pi}_\alpha = \hat{F}_n^{-1}(\alpha)$ from the data: $\tilde{\text{VaR}} = -\hat{F}_n^{-1}(\alpha)$
  - OK if α is not too small relative to sample size n
- Semiparametric approach with threshold u: note that

$$F(x) = P(X \leq x) = P(X \leq x | X \leq u) P(X \leq u)$$

  - Utilize POT estimate the parametric tail probability model of F_θ for $F(X \leq x | X \leq u)$ (e.g., GPD), i.e.,
  - Utilize nonparametric estimate of $P(X \leq u)$ -
  - Estimate quantile from above via

$$\tilde{\text{VaR}} = -\hat{F}_{spar}^{-1}(\alpha) = \qquad .$$

Computing VaR for SP500 Weekly Returns

Example. Based on results in previous slides on fitting GPD to the tails, want to derive VaR for $\alpha = .005$ and current investment value of a million dollars.

Answer. .

R-code for Answer

```
> eecdf = ecdf(WSPLRet)
> alphat = 1-.005/eecdf(-.02)
> scale = gpd_fit$fitted.values[1]
> xi = gpd_fit$fitted.values[2]
> xi
0.1225712
> m = .02
> VaRt = qgpd(alphat,m,scale,xi)
> VaRt
0.06759757
> VaR = 1000000*VaRt
> VaR
67,597.57
```

Assumptions/Issues

Question. What are assumptions/issues for the utilizing tail analysis as proposed?