Describing and Comparing Continuous Univariate Distributions: Brief Introduction

By

Daniel B. Carr

George Mason University

Common Descriptors, Moments, and Some R Functions

Support: domain of positive density

Sample size: n = length(x) # x a vector or data

Location: often related to $\mu = E(X)$

Mean: xMean = mean(x)

Standard error of the mean: xMeanSe = sqrt(var(x))/n

95% confidence interval: xMean ± qt(.975,df=30)*xMeanSe

Median: x such that Pr[X ≤ x] == Pr[X ≥ x] == .5

Mode: locations of local maxima

Scale or spread: often related to $E[(X-\mu)^2]$

Standard deviation:	sqrt(var(x))
Interquartile range:	3rd - 1st quartiles or .75 quantile - .25 quantile
Range (max - min)	range(x) == c(min(x), max(x))
	diff(range(x)) == max(x) - min(x)

Higher order moments

Skewness related to $E[(X-\mu)^3]$

Kurtosis or tail thickness related to $E[(X-\mu)^4]$

Distributional Caricatures or Schema

Mean and 95% confidence interval

Box plots: a five number summary plus outliers

Median

First and third quartiles

Lower adjacent value

Smallest observed value > 1^{st} quartile – 1.5 * Interquartile range

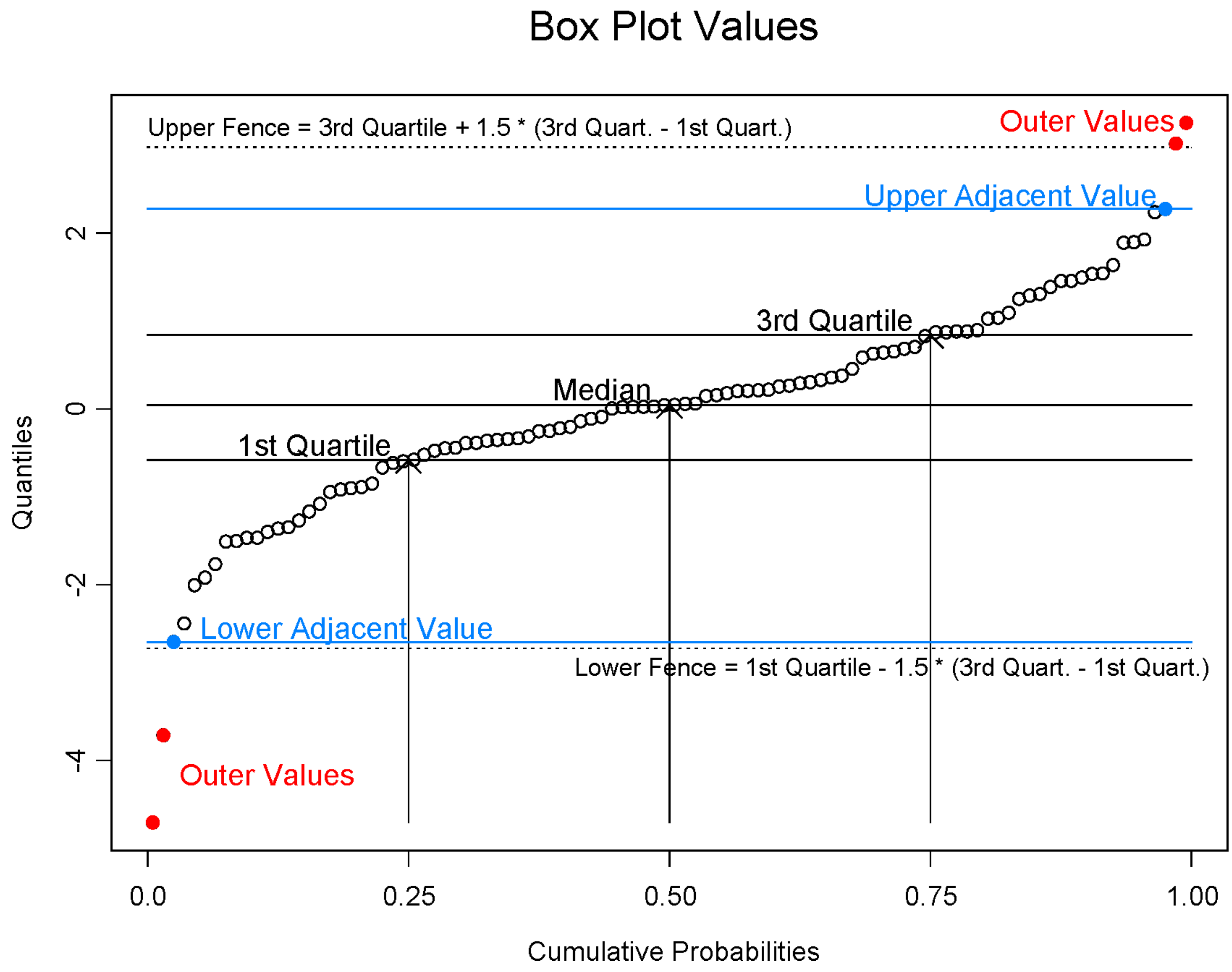
Upper adjacent value

Largest observed value < 3^{rd} quartile + 1.5 * Interquartile range

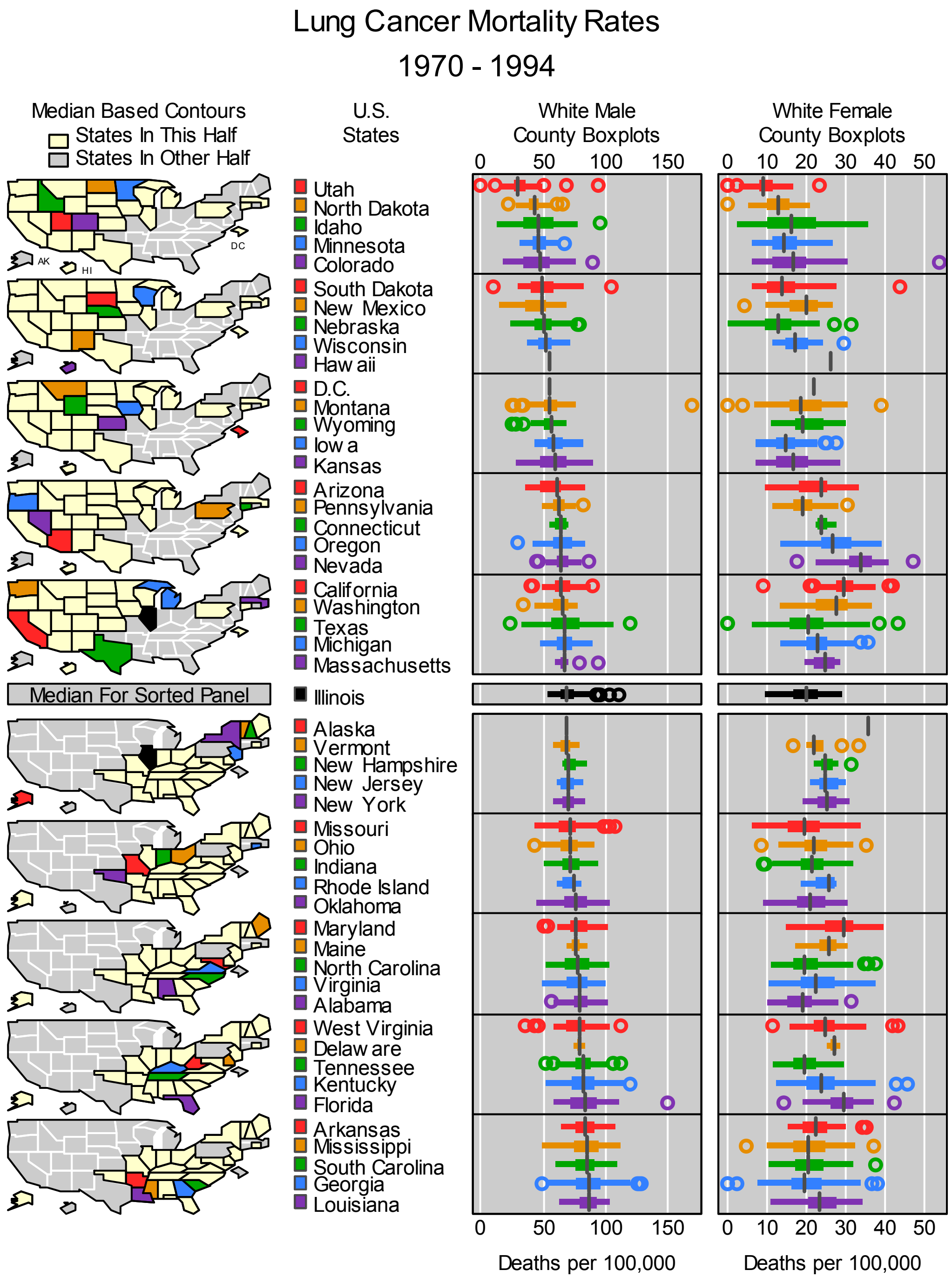
Outliers

Notes: Sometimes adjacent values are the min and the max

Some box plots variations do not compute adjacent values and just use the minimum and the maximum



Box Plot Values
Upper Fence = 3rd Quartile + 1.5 * (3rd Quart. - 1st Quart.)
Outer Values
Upper Adjacent Value
3rd Quartile
Median
1st Quartile
Lower Adjacent Value
Lower Fence = 1st Quartile - 1.5 * (3rd Quart. - 1st Quart.)
Outer Values
Quantiles
2
0
-2
-4
0.0
0.25
0.50
0.75
1.00
Cumulative Probabilities


Lung Cancer Mortality Rates
1970 - 1994
Median Based Contours
States In This Half
States In Other Half
U.S.
States
White Male
County Boxplots
White Female
County Boxplots
0
50
100
150
0
10
20
30
40
50
DC
AK
HI
Utah
North Dakota
Idaho
Minnesota
Colorado
South Dakota
New Mexico
Nebraska
Wisconsin
Hawaii
D.C.
Montana
Wyoming
Iowa
Kansas
Arizona
Pennsylvania
Connecticut
Oregon
Nevada
California
Washington
Texas
Michigan
Massachusetts
Median For Sorted Panel
Illinois
Alaska
Vermont
New Hampshire
New Jersey
New York
Missouri
Ohio
Indiana
Rhode Island
Oklahoma
Maryland
Maine
North Carolina
Virginia
Alabama
West Virginia
Delaware
Tennessee
Kentucky
Florida
Arkansas
Mississippi
South Carolina
Georgia
Louisiana
Deaths per 100,000
Deaths per 100,000

Box Plot Styles and Options

Common style

```
boxplot(count ~ spray, data = InsectSprays,
        horizontal=T, notch=F, col = "lightgray", las=1,
        xlab="Number of Insects", ylab="Type of Spray",
        main= "R-Style Boxplots")
```

Notches if chose provide comparison intervals the medians

Trellis variations and options

Easily seen dot for the median

Interquartile range rectangle width proportional to the count

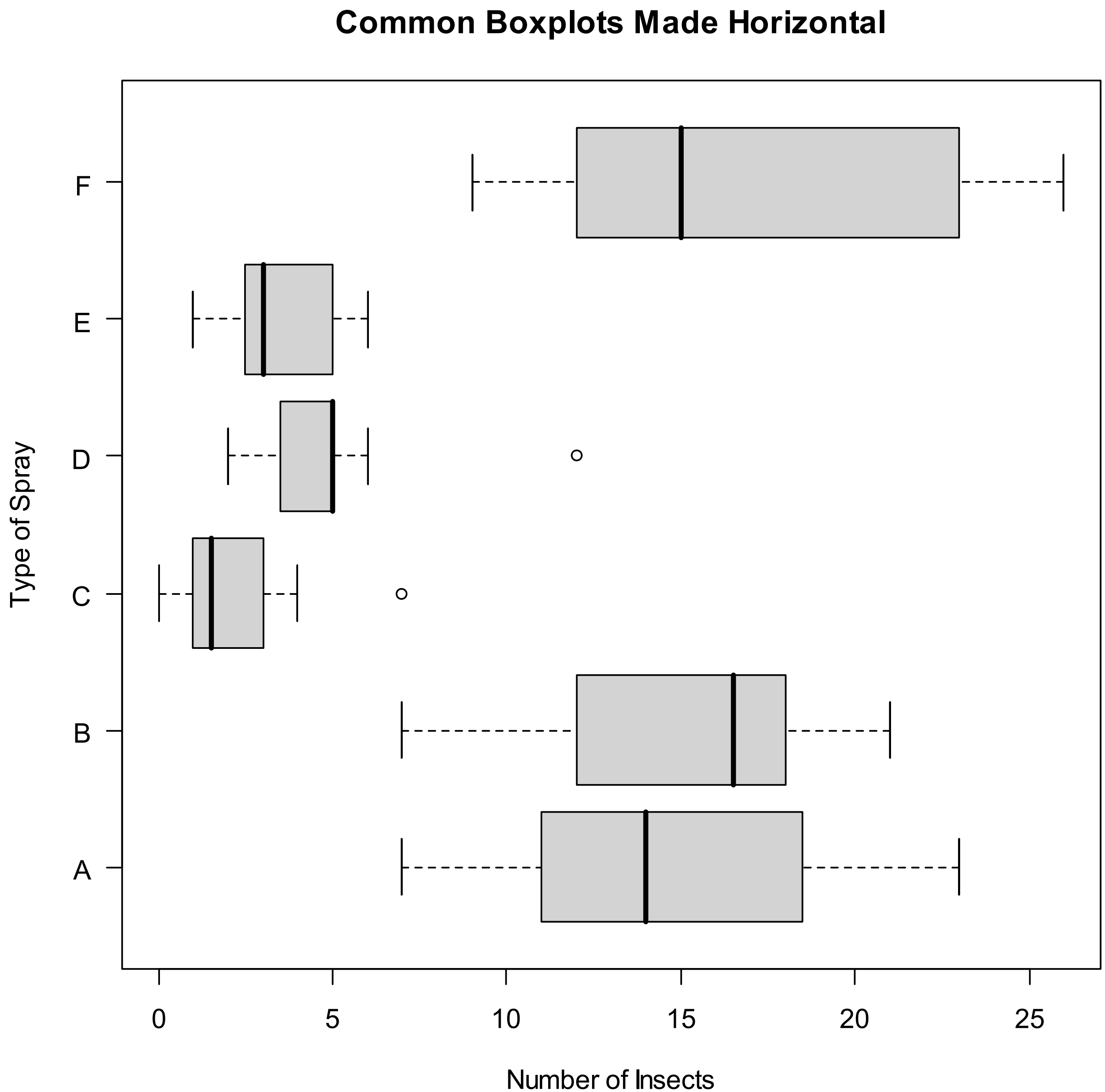
Instructor's style – see newsletter article

Median line extends outside the interquartile range rectangle

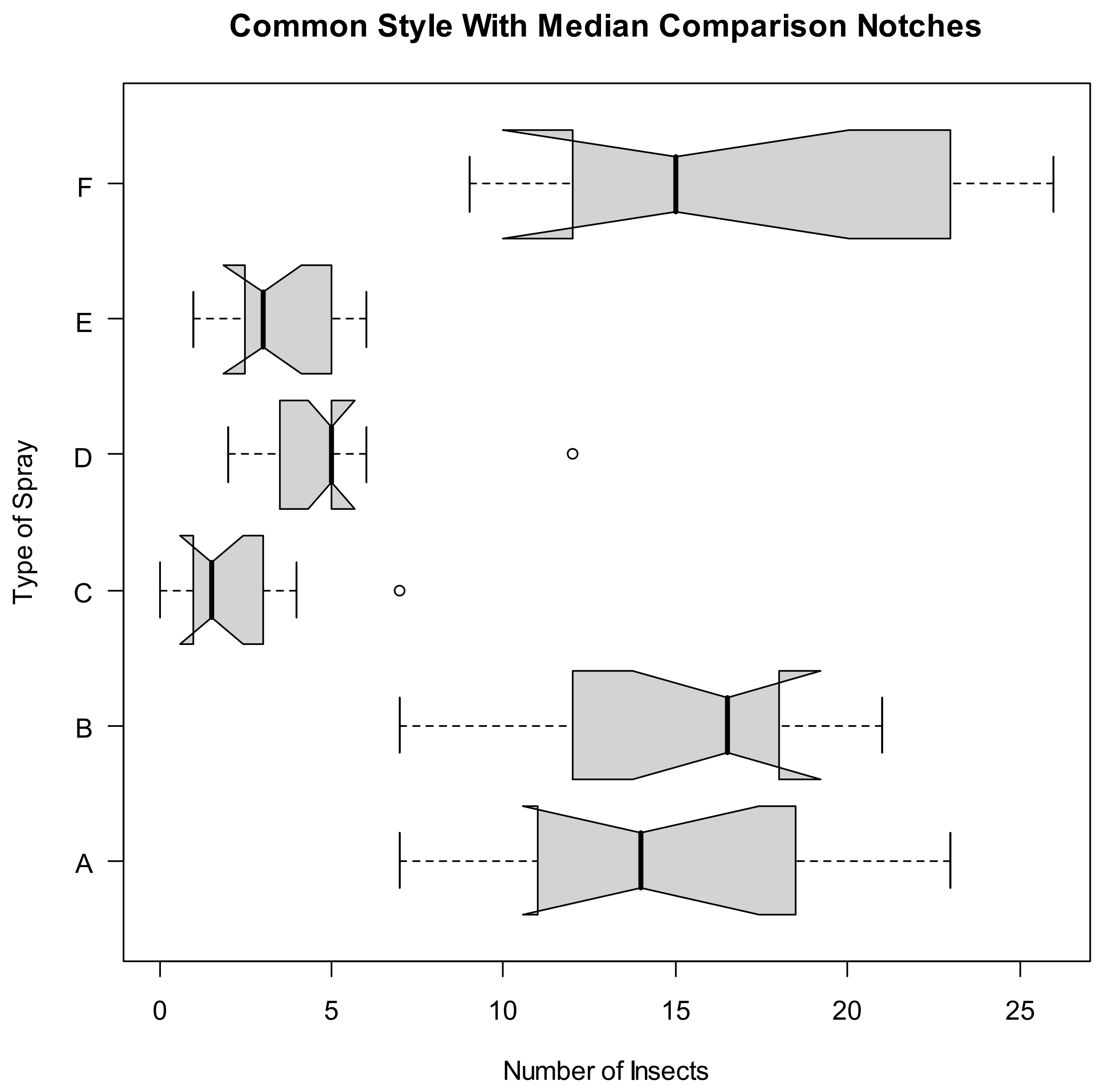
Thinner rectangles from interquartile range rectangle to adjacent values

Notch interval (if shown)

white line on colored rectangles but under the median line



Common Boxplots Made Horizontal
F
E
D
C
B
A
Type of Spray
0
5
10
15
20
25
Number of Insects


Common Style With Median Comparison Notches
Type of Spray
F
E
D
C
B
A
0
5
10
15
20
25
Number of Insects

More Detailed Distributional Descriptions

Density functions f(x) for continuous random variable

$f(x) \geq 0$

$\int f(x)\, dx = 1$

Normal (or Gaussian) distribution: 2 parameter family

Support: $(-\infty, \infty)$

Mean μ: location parameter

Standard deviation σ, spread parameter

Commonly use in distributional comparisons

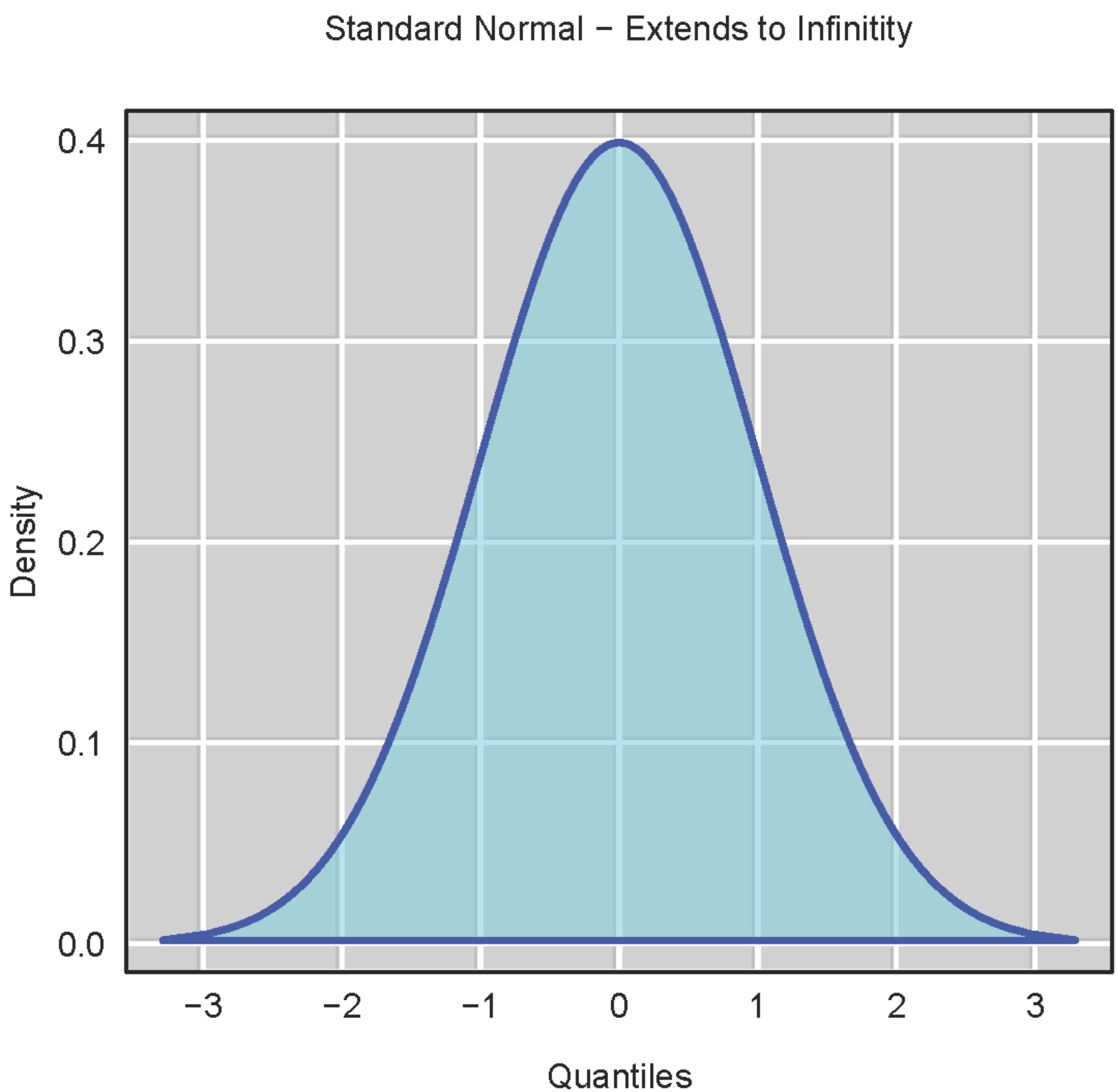
Tail thickness: $con*exp(-x^2)$

Multivariate extension a special case:

Correlation (x, y) = 0 implies independence

Not true in general

Familiar bell-shaped curve


Standard Normal − Extends to Infinitity
0.4
0.3
0.2
0.1
0.0
Density
−3
−2
−1
0
1
2
3
Quantiles

Theoretic Density Plot Production

```
source ('gridPlotFunctions.R')

n = 1000
cumprobs = (1:n-.5)/n
quants = qnorm( cumprobs )
normDensity = dnorm( quants )
windows()
gPlot( quants, normDensity, type='l', lwd=3, col="blue",
      xlab='Quantiles', ylab='Density', main="Standard Normal")

#Alphablending fill version

pdf( "Standard Normal.pdf", version="1.4")
gPlot( quants, normDensity, type='n',
xlab='Quantiles', ylab='Density', main="Standard Normal - Extends to Infinitity")
polygon( quants, normDensity, col="#00FFFF88", border="blue", lwd=3)
dev.off()
```

Adobe Acrobat Profession will convert pdf to .png or .jpeg files

Density Estimation: Kernel Approach

Steps

Treat each observation as partial evidence for its neighborhood

Center a local likelihood bump (density) at each observation

Add up the bumps

Divide by the number of observations to produce an average

Notes

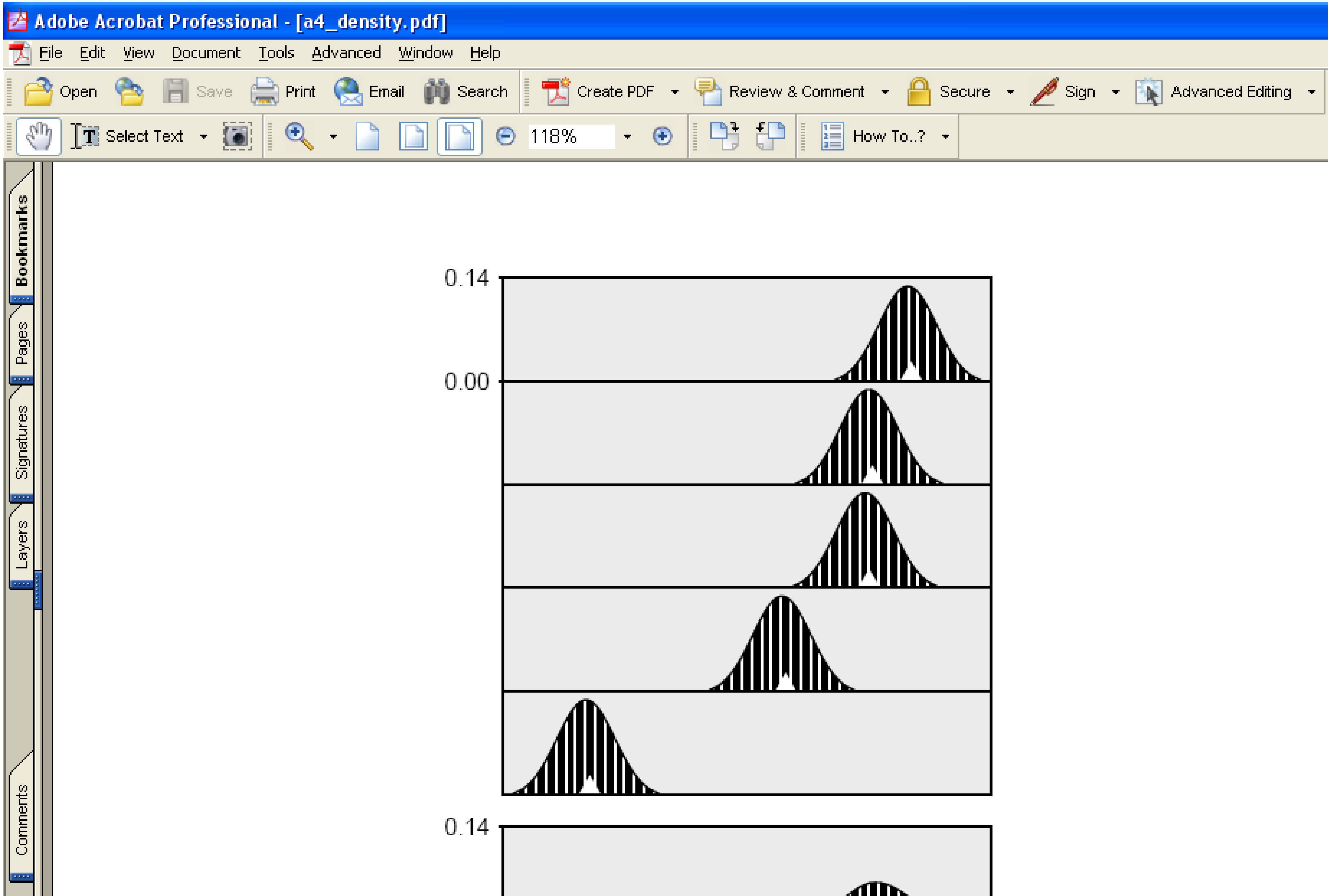
Smoothing parameter: the spread of the bumps

Selection can be based on theory and a cross validation

Bumps are densities so the average is a density

The derivative of the density estimate is related directly to the derivatives of the bumps

People tend to understand density plots more readily than quantile and cumulative probability plots


Adobe Acrobat Professional - [a4_density.pdf]
File Edit View Document Tools Advanced Window Help
Open
Save
Print
Email
Search
Create PDF
Review & Comment
Secure
Sign
Advanced Editing
Select Text
118%
How To..?
Bookmarks
Pages
Signatures
Layers
Comments
0.14
0.00
0.14

Density Estimation Approaches

Non-Parametric

Kernel estimation and other methods

Parametric

Chose a parametric family:

Normal, t, gamma, beta, weibull, …

Estimate the parameters via maximum likelihood

Semi-parametric

Example: Estimate parameters for a mixture of normals

Empirical Quantile and Cumulative Probability Plots

Quantile plots: y= quantiles, x= cumulative probabilities

Cumulative probability plots: y= cumulative probabilities, x = quantiles

Empirical quantiles are sorted values or order statistics

```
dat = c ( 7.2, 3.5, 9.1, 12.4, 3.0 )
quants = sort ( dat )
```

Formula for cumulative probabilities in this class

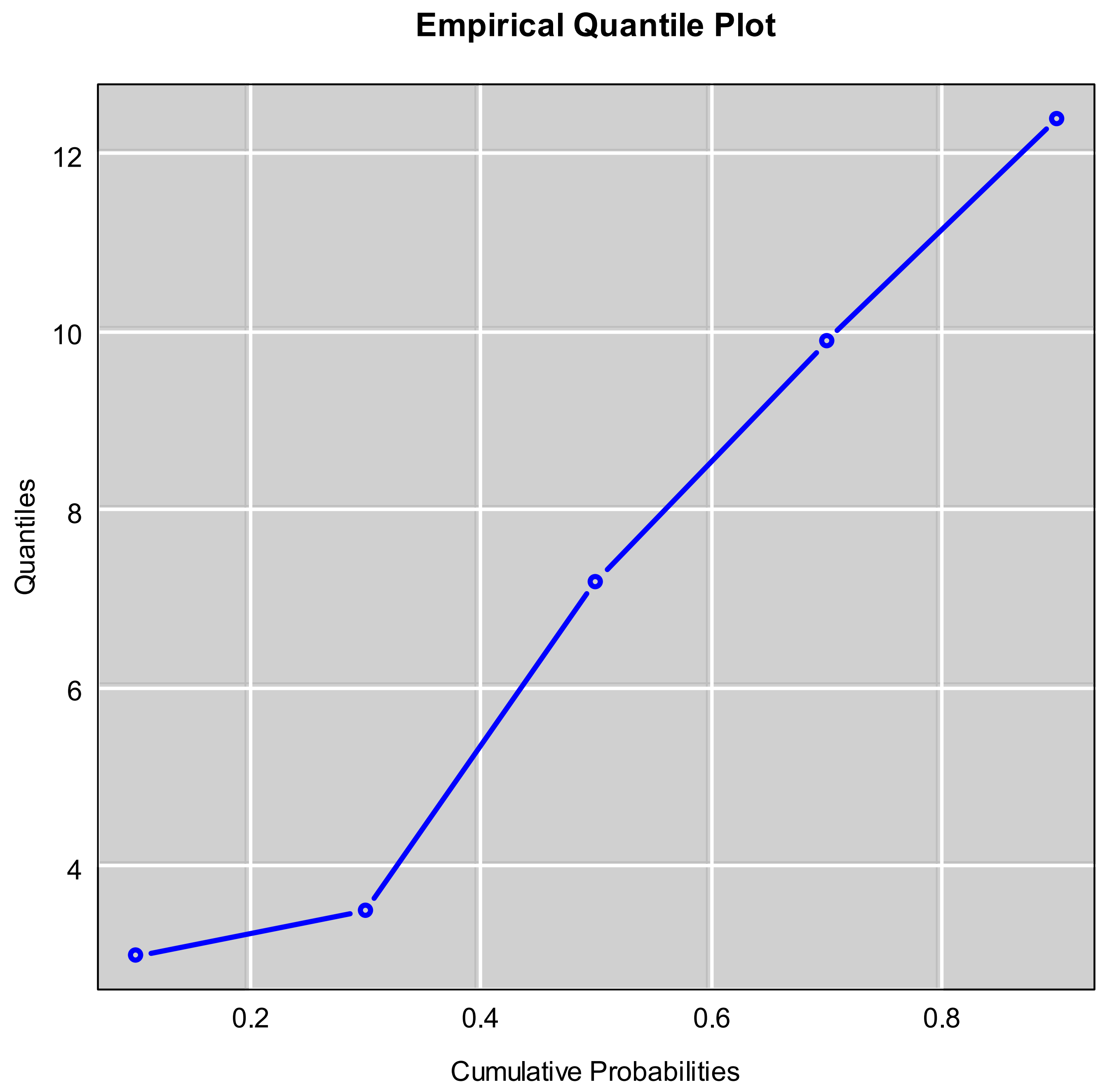
n = length(dat)

cumProbs = (1:n - .5)/n # i^{th} value = (i - 5) / n

Other approach include empirical cumulative probabilities

Paired values: cbind(cumProbs, quants)

cumProbs	Quantiles		
0.1	3.0	#	0.1 = (1 - .5) / 5
0.3	3.5	#	0.3 = (2 - .5) / 5
0.5	7.2	#	0.5 = (3 - .5) / 5
0.7	9.1	#	0.7 = (4 - .5) / 5
0.9	12.4	#	0.9 = (5 - .5) / 5



Empirical Quantile Plot
Quantiles
12
10
8
6
4
0.2
0.4
0.6
0.8
Cumulative Probabilities

R Distribution Functions

R has distribution functions for many continuous parametric families

Base family names include

beta, cauchy, chisq, exp, f, gamma, lnorm, norm, t, unif, weibull

Prefixes determine the kind of values returned

p stands for cumulative probabilities
q stands for quantiles
d stands for density values (ordinates)
r stand for random number from

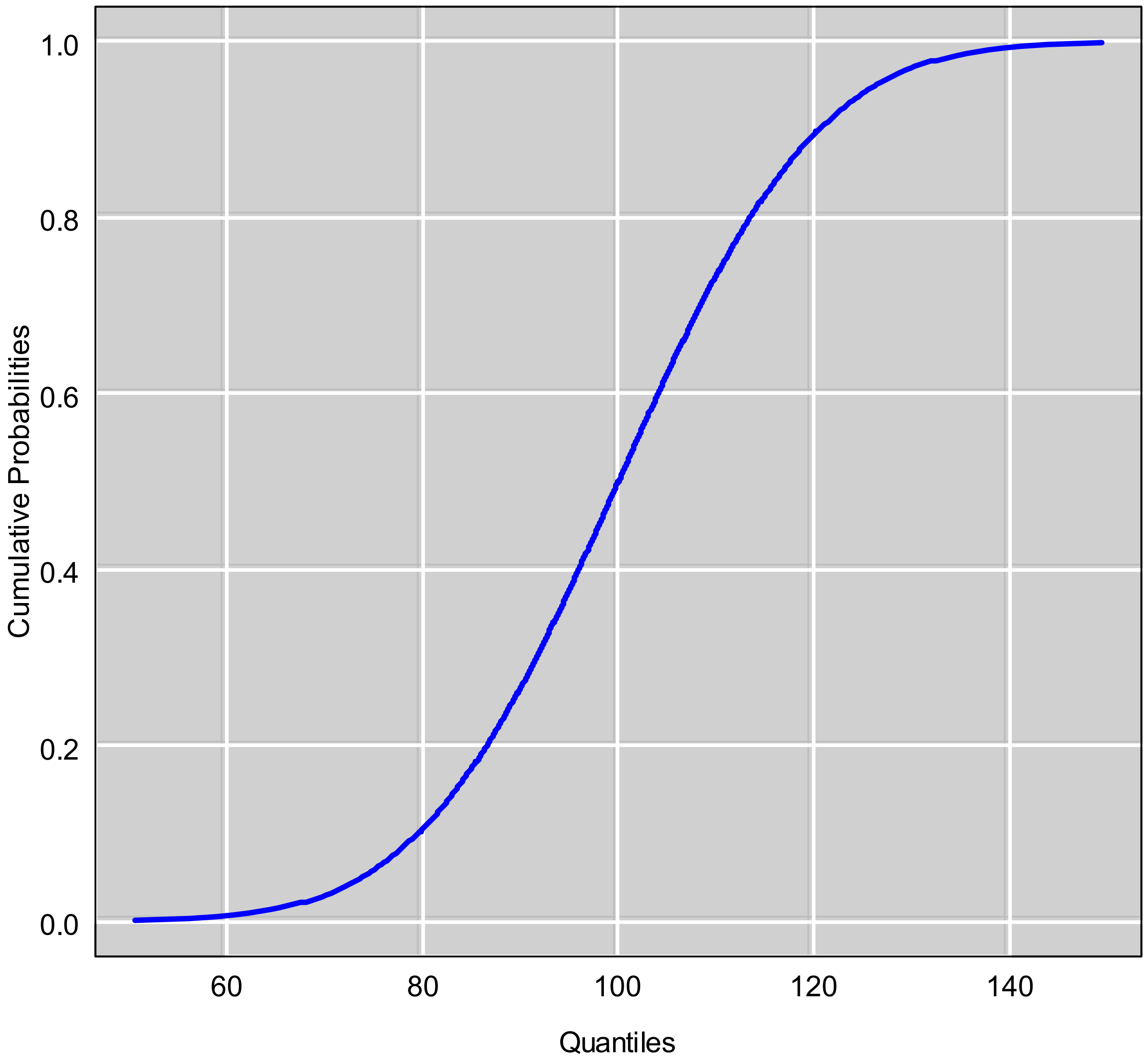
Primary arguments depend on prefix

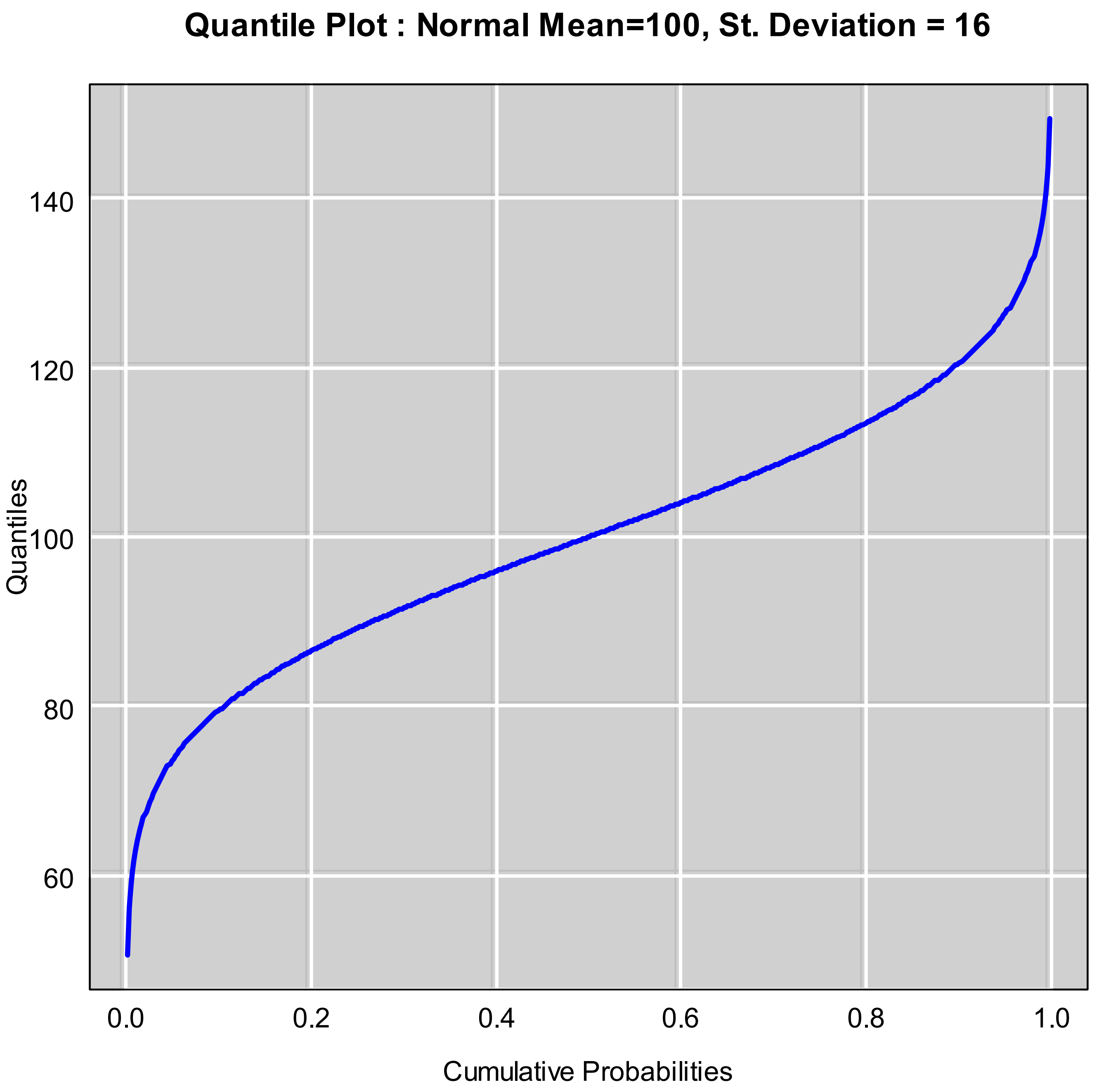
quantiles for cumulative probabilities and densities
cumulative probabilities for quantiles
sample size for random numbers

Defaulted secondary arguments indicate family specify parameters

```
qnorm(cumprobs, mean=100, sd=16) # all arguments can be vectors!
pt (quants, df=17)               # all argumetns can be vectors
runif(100, min=-1, max=1)
```

Cumulative Probabilty Plot: Normal Mean=100, St. Deviation = 16




Quantile Plot : Normal Mean=100, St. Deviation = 16
Quantiles
140
120
100
80
60
0.0
0.2
0.4
0.6
0.8
1.0
Cumulative Probabilities

Superimposed Quantile Plot Comparison

Visual comparison is not so easy

We don't see differences in the technically correct direction

Interpretation is not easy when there are scale difference

This is common in comparing different empirical distributions

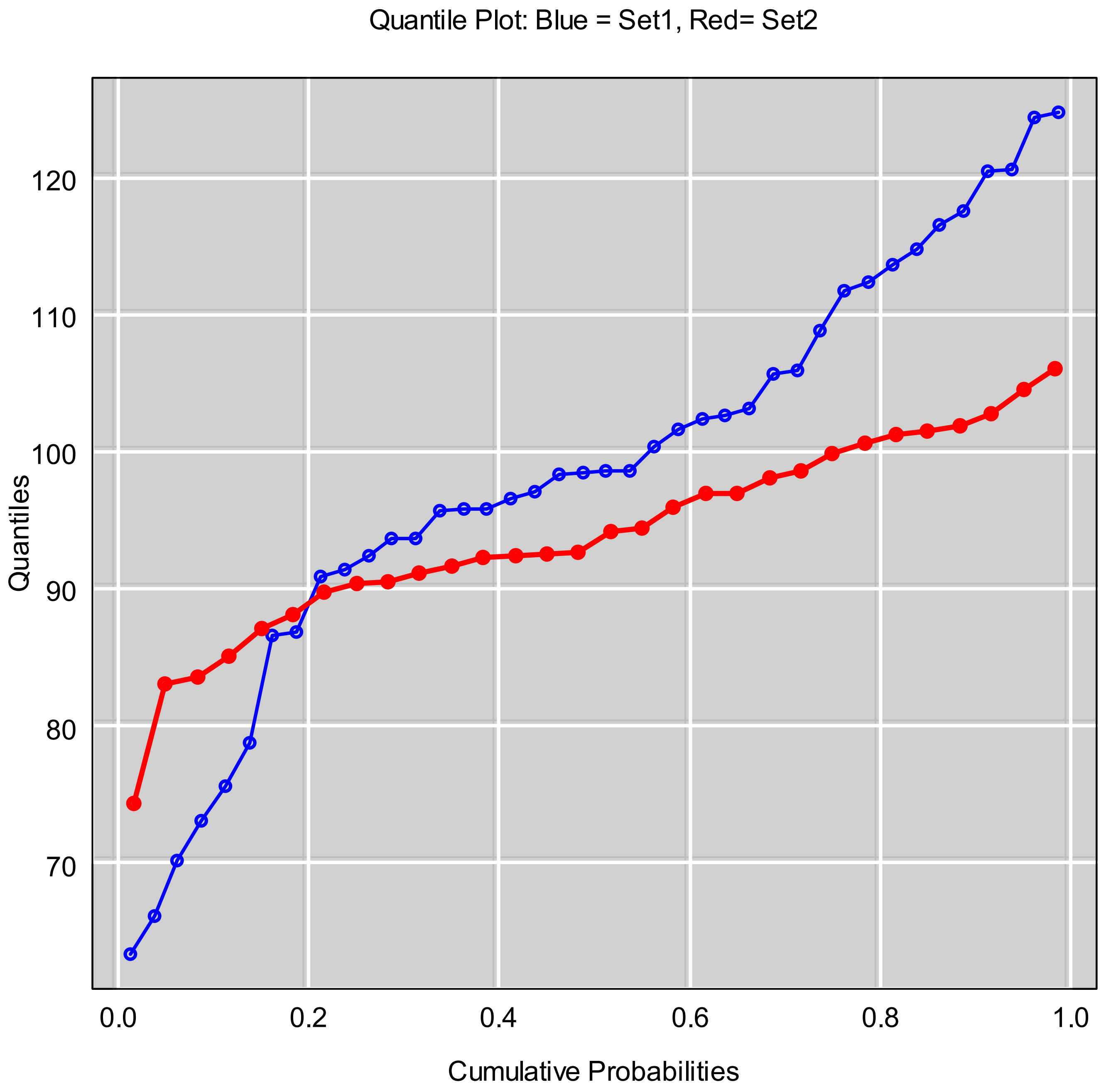
Two approaches to matched comparisons

PP plots: Match quantiles and plot paired cumulative probabilities

QQ plots: Match cumulative probabilities and plot paired quantiles

Class preference: QQ plots

If the distributions differ only in location and scale, they will appear as a straight line. (There will be noise in samples.)



Quantile Plot: Blue = Set1, Red= Set2
Quantiles
120
110
100
90
80
70
0.0
0.2
0.4
0.6
0.8
1.0
Cumulative Probabilities

QQplots

Interpolation to obtain cumulative probability matched quantiles

The quantile() function produces interpolated quantiles for given cumulative probabilities
Using identical cumulative probabilities provides matched quantiles
Restriction: The smaller data set limits the range of common cumulative probabilities

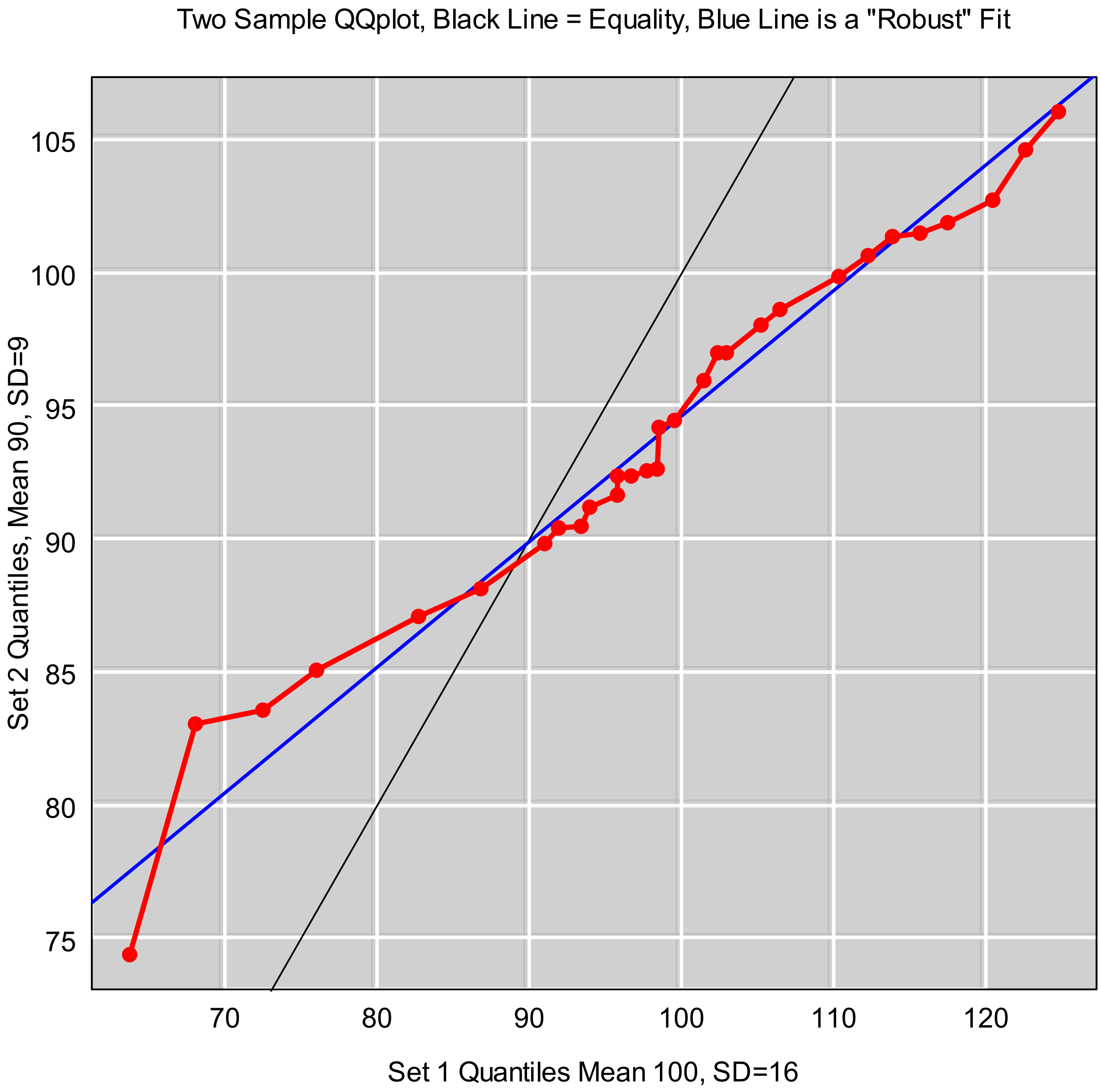
Normal QQplots

Comparison of empirical quantiles (y) against standard normal quantiles (x) is common.
Multiple sets of empirical quantiles can be plotted versus the same globally scaled x-axis.
For a single set, normal quantiles can have the same mean and deviation as the set
This changes the x-axis scale but not the plot appearance

We are often looking for skewness and departures in tail thickness
When the empirical values are regression residuals such departures are warning flags

Visual Aids

Fitting a robust fit straight line
Showing the intercept 0 slope 1 reference line
Transforming to show Mean (x-axis) and Differences (y-axis) in a MD plot


Two Sample QQplot, Black Line = Equality, Blue Line is a "Robust" Fit
105
100
95
90
85
80
75
70
80
90
100
110
120
Set 2 Quantiles, Mean 90, SD=9
Set 1 Quantiles Mean 100, SD=16