Data Analysis for Mechanical Engineering

General Principles/Philosophy

William 'Ike' Eisenhauer

Department of Mechanical and Materials Engineering
Portland State University
Portland, Oregon 97223

wde@pdx.edu

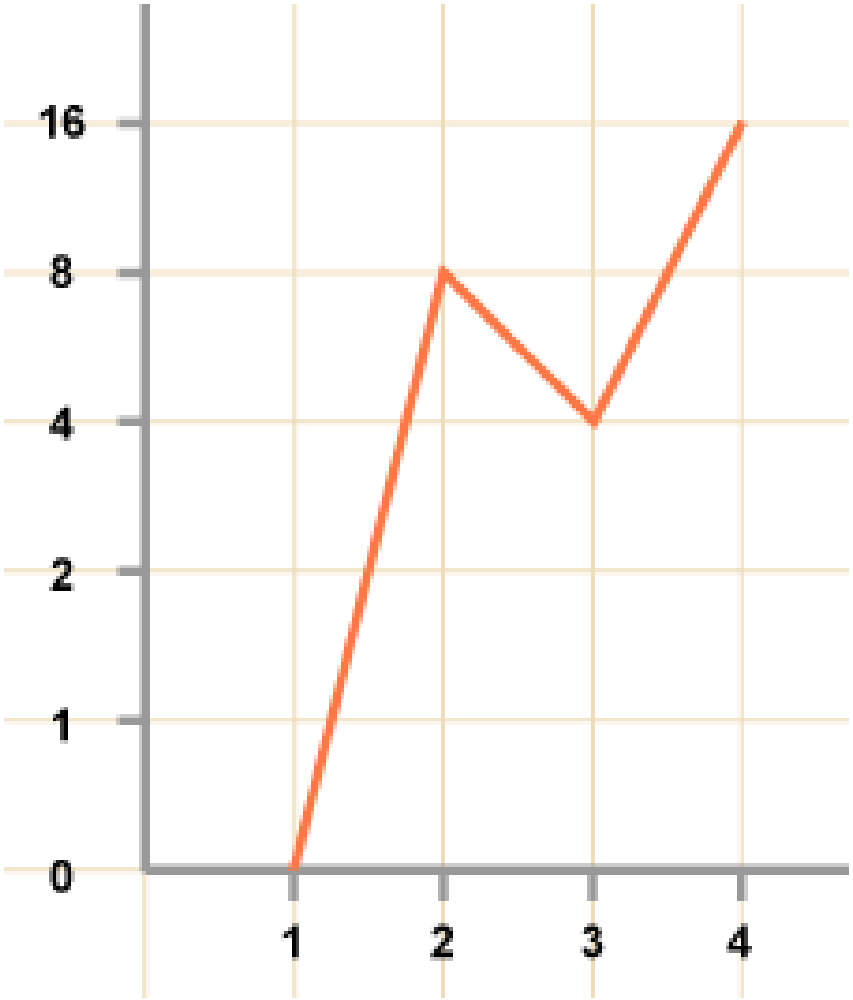
Winter 2016

Bad Graphs

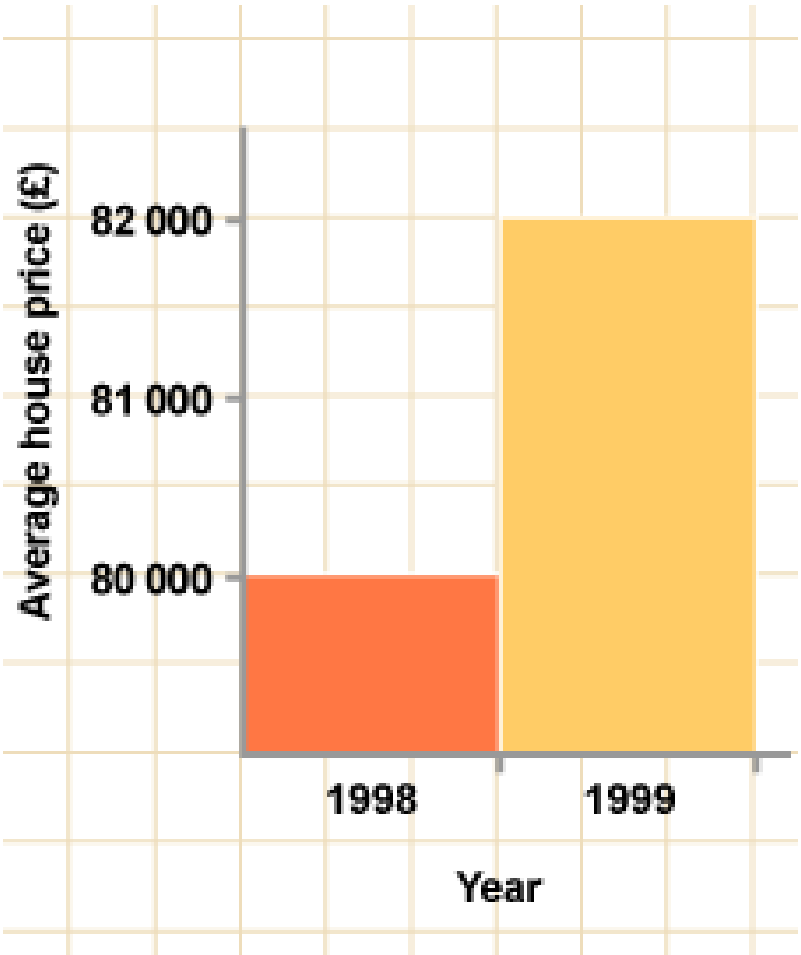
Sometimes...

- Graphs are used to get emotion
- Graphs are used to confuse
- Graphs are used for snazzy presentations
- Graphs are used wrong...

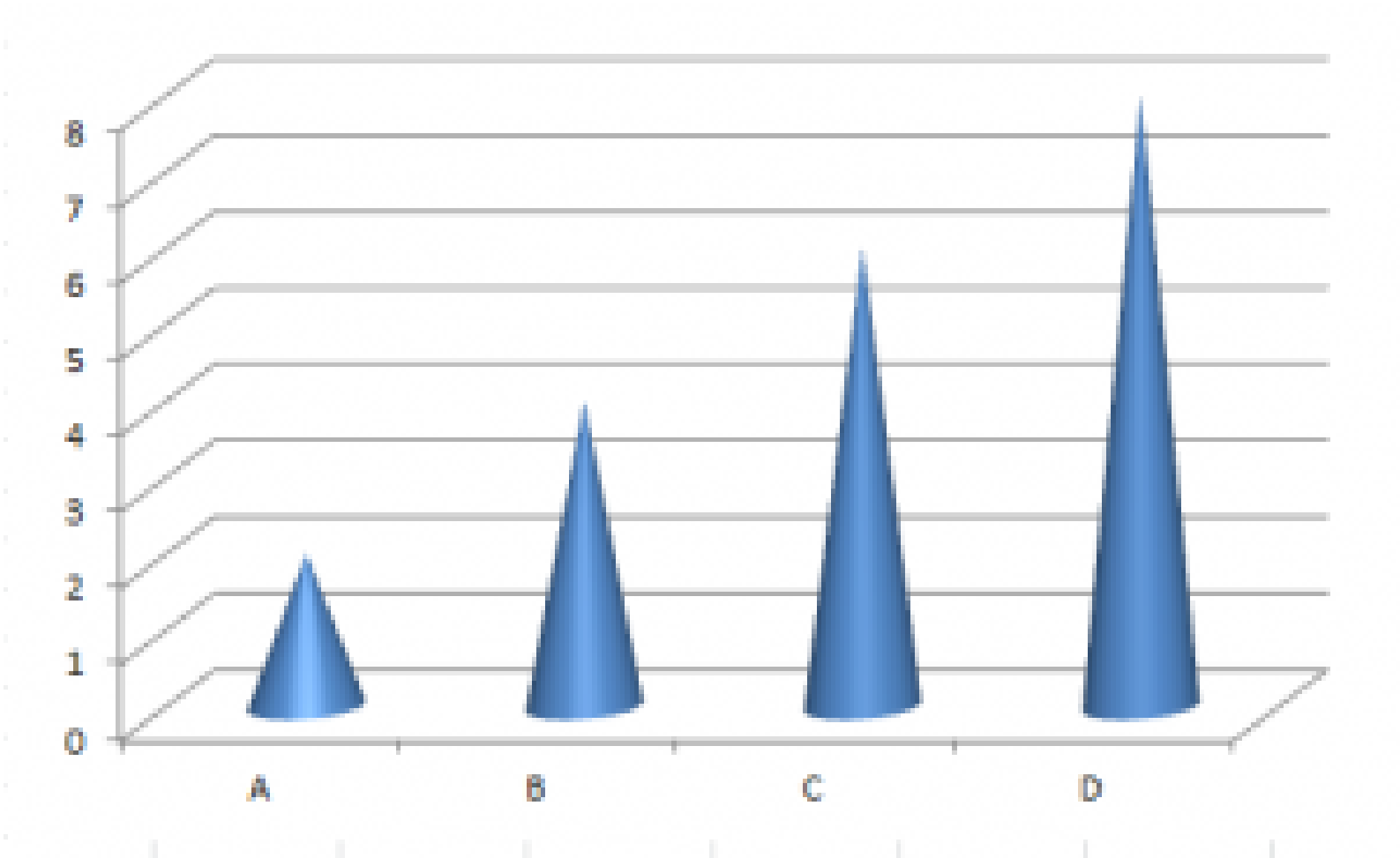
What's wrong...



What's wrong...



What's wrong...



What's wrong...

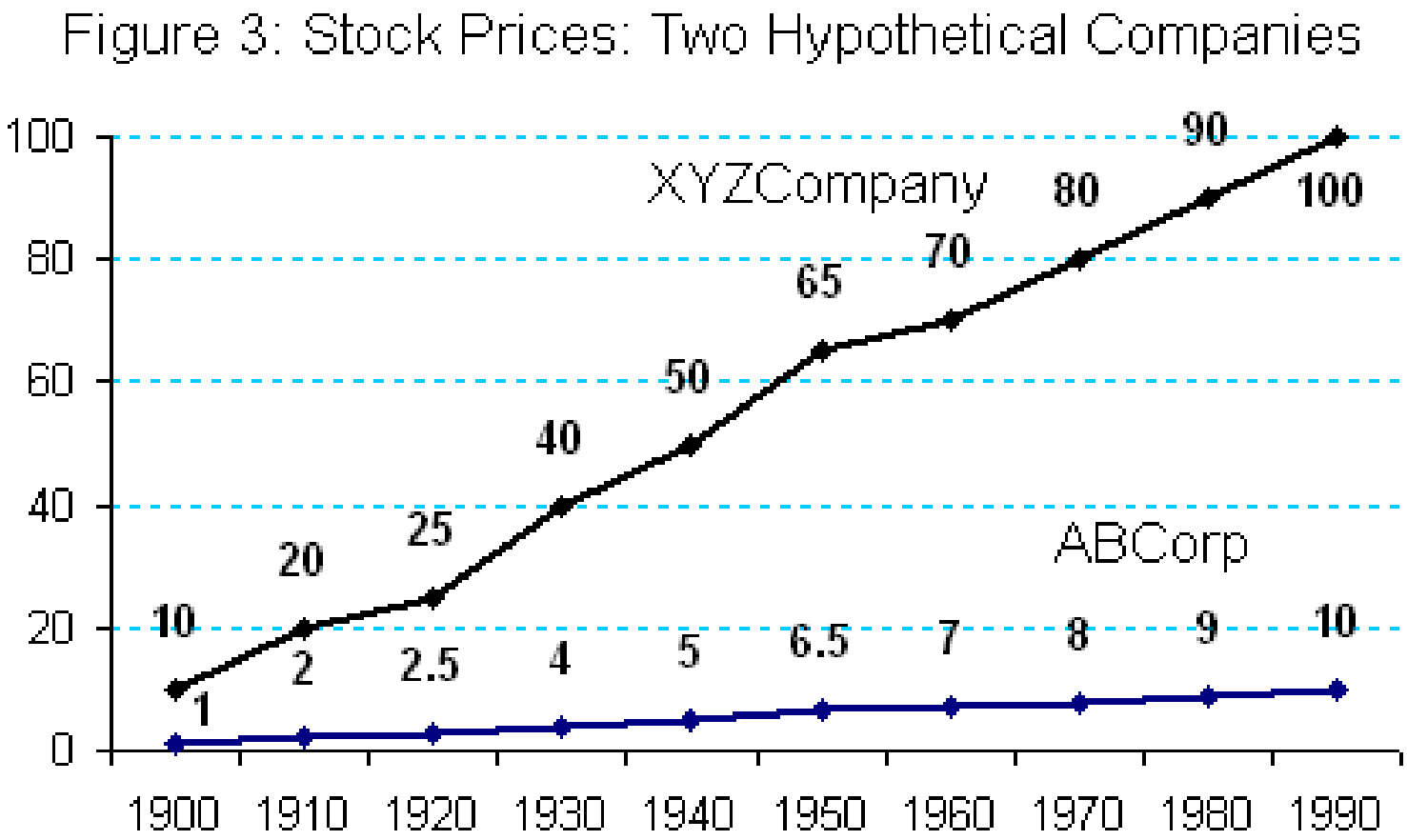
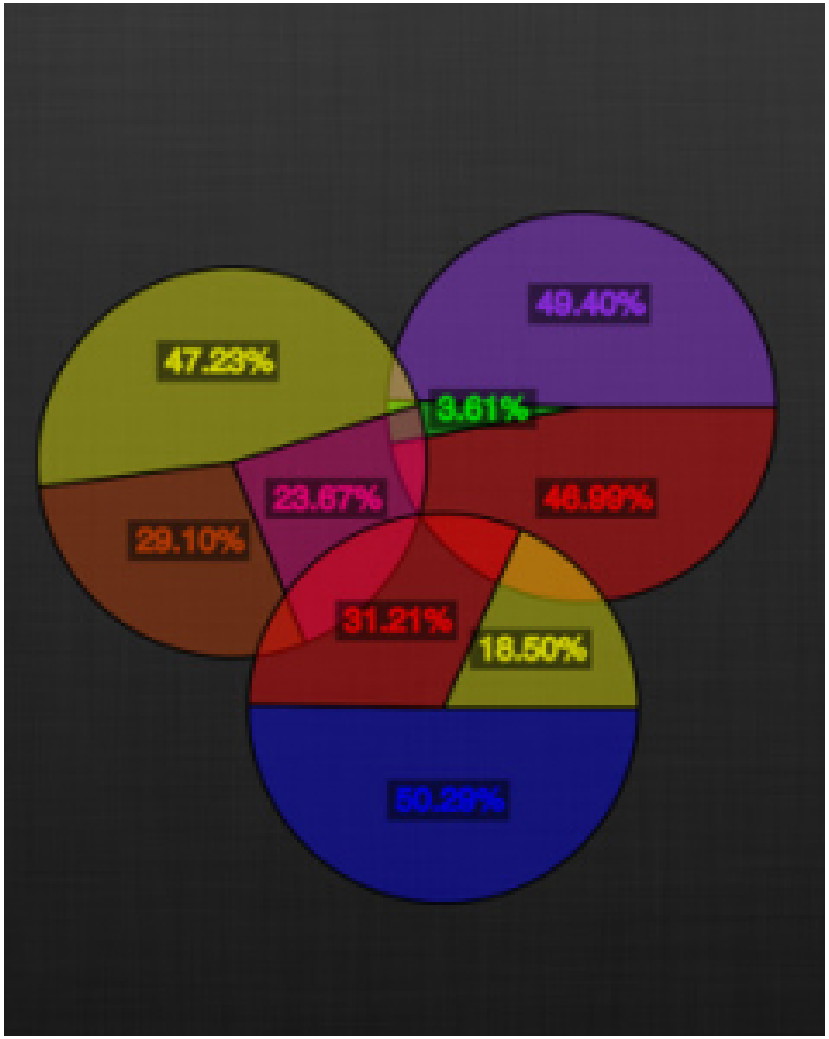


Figure 3: Stock Prices: Two Hypothetical Companies
100
80
60
40
20
0
XYZCompany
ABCorp
10
20
25
40
50
65
70
80
90
100
1
2
2.5
4
5
6.5
7
8
9
10
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990

What's wrong...



What's wrong...

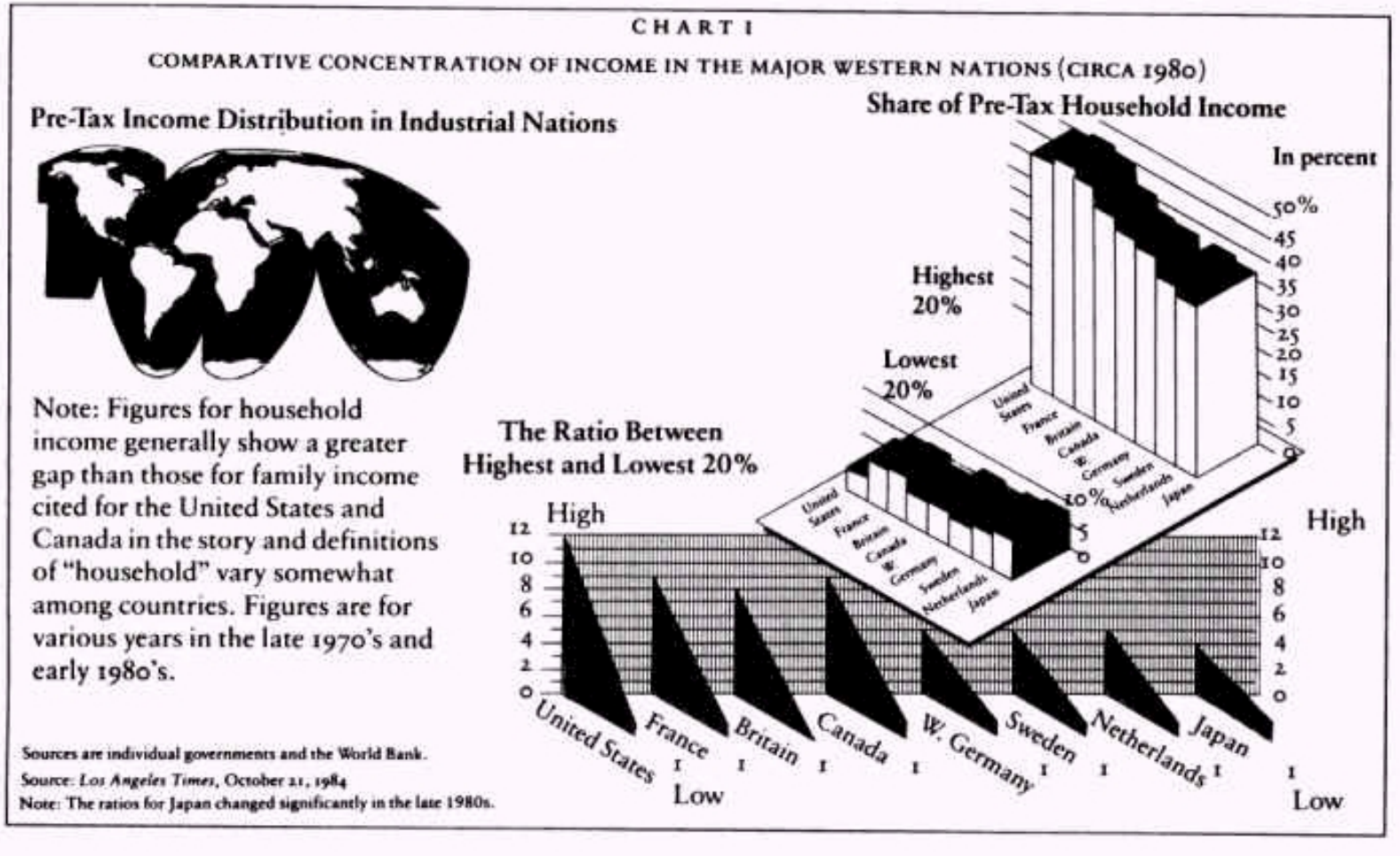
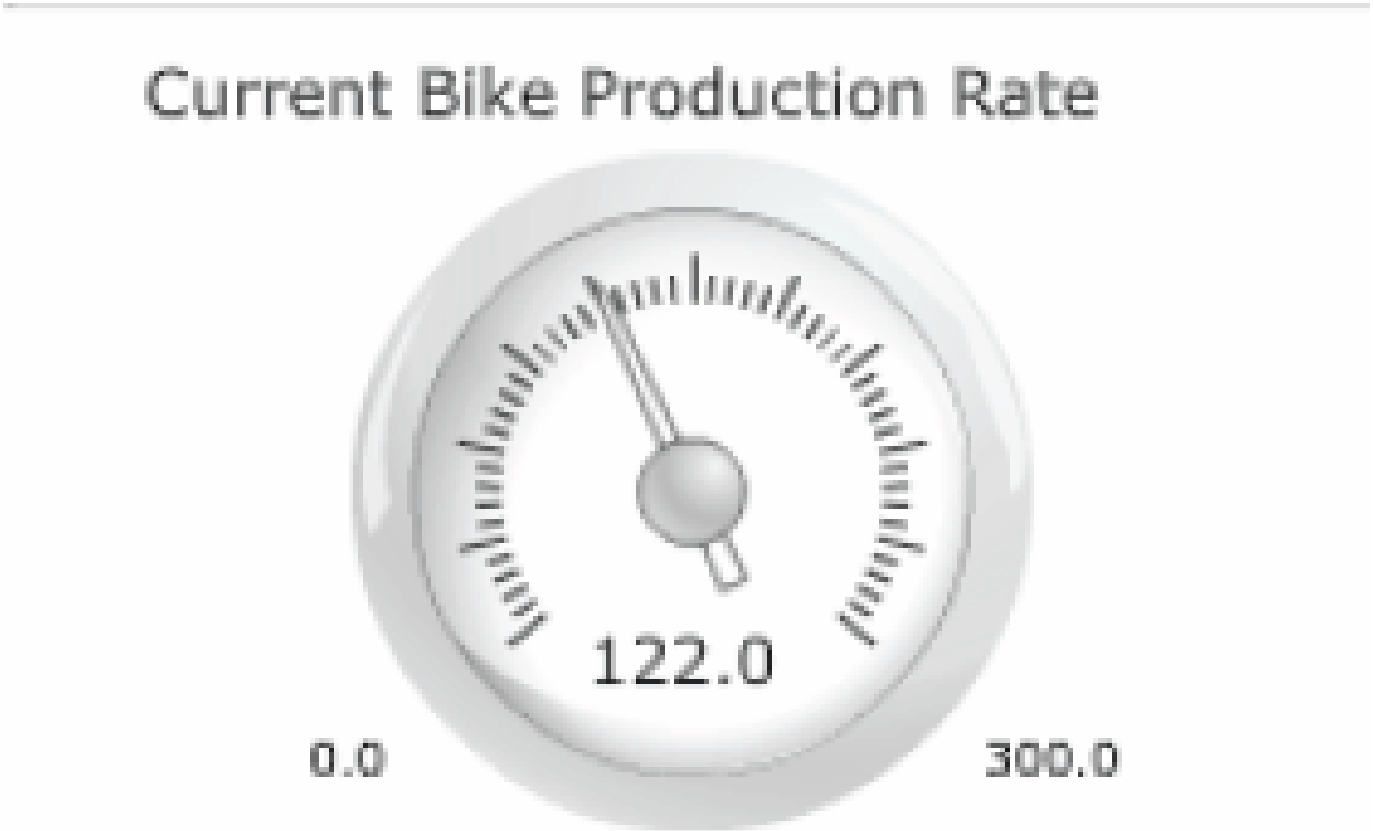
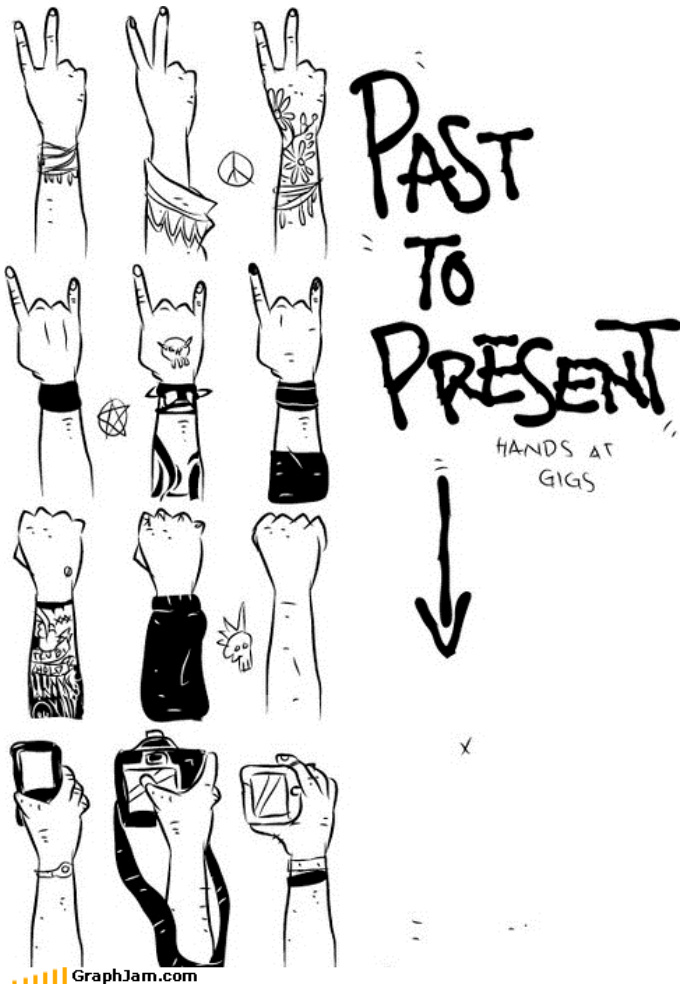


CHART 1
COMPARATIVE CONCENTRATION OF INCOME IN THE MAJOR WESTERN NATIONS (CIRCA 1980)
Pre-Tax Income Distribution in Industrial Nations
Note: Figures for household income generally show a greater gap than those for family income cited for the United States and Canada in the story and definitions of "household" vary somewhat among countries. Figures are for various years in the late 1970's and early 1980's.
Sources are individual governments and the World Bank.
Source: Los Angeles Times, October 21, 1984
Note: The ratios for Japan changed significantly in the late 1980s.
Share of Pre-Tax Household Income
In percent
Highest 20%
Lowest 20%
The Ratio Between Highest and Lowest 20%
High
Low
United States
France
Britain
Canada
W. Germany
Sweden
Netherlands
Japan

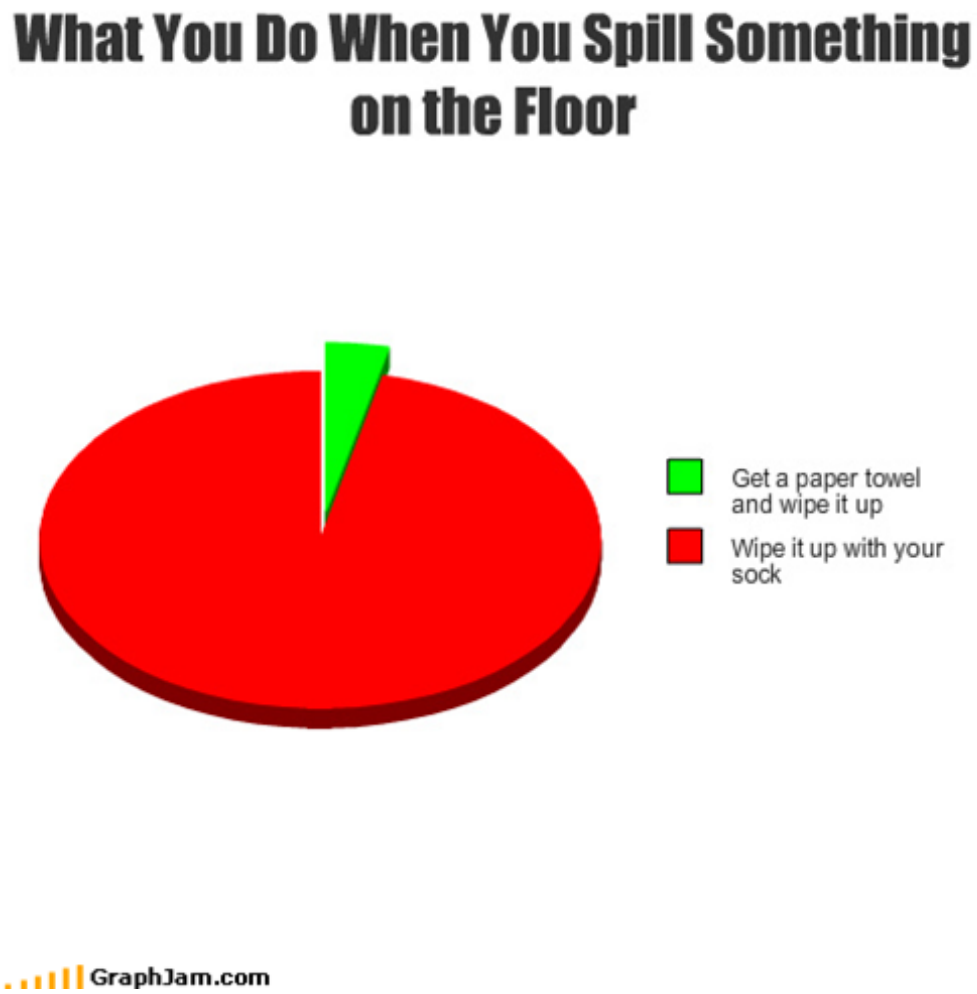
What's wrong...



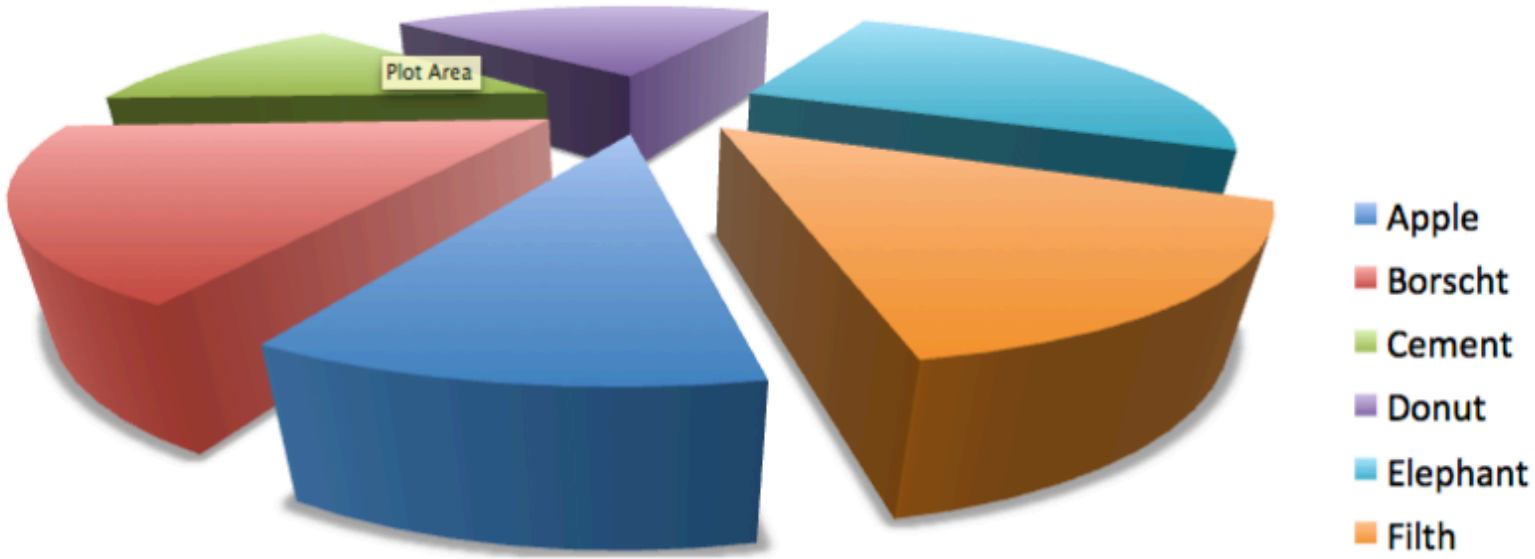
What's wrong...



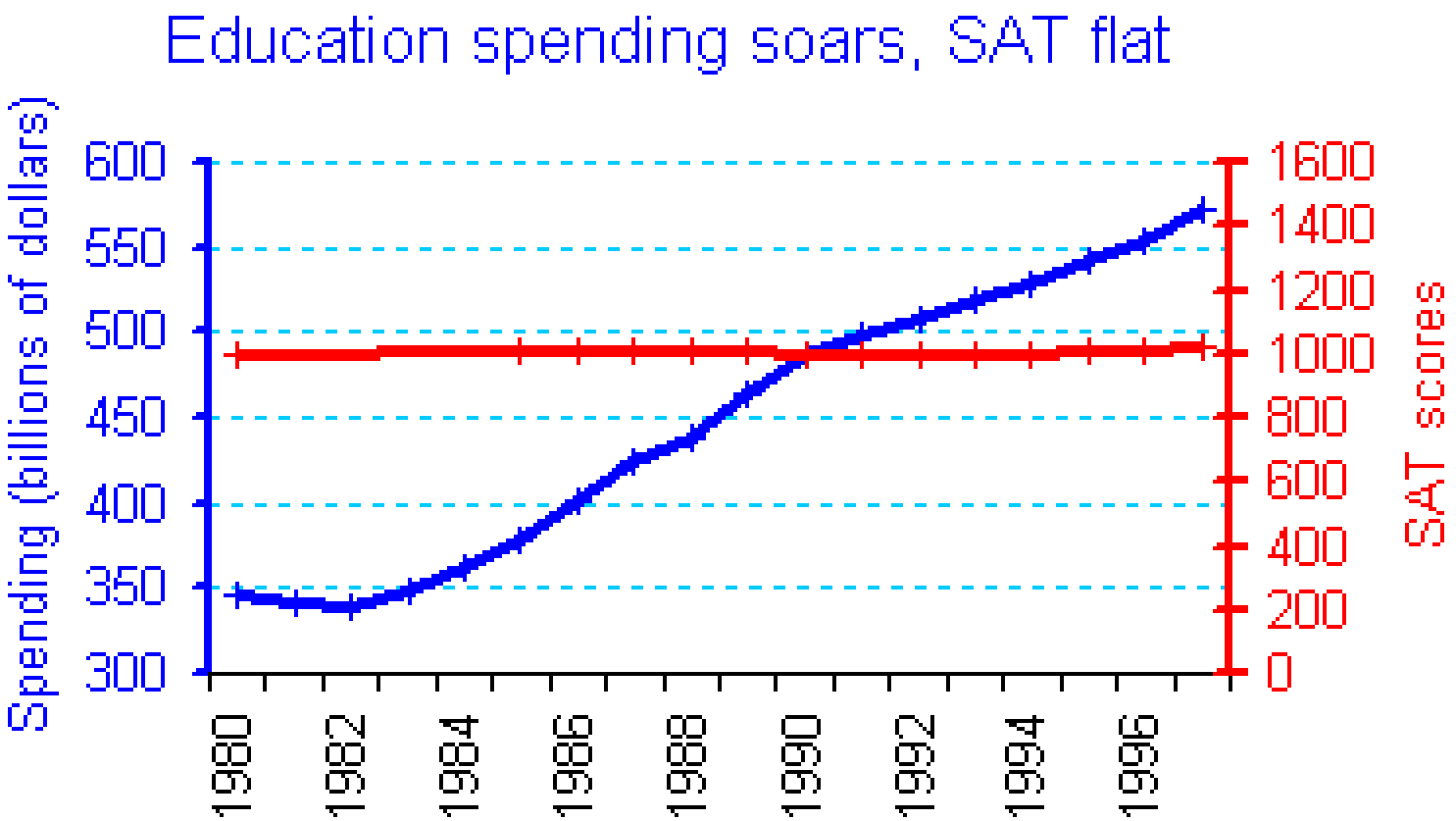
What's wrong...


What You Do When You Spill Something on the Floor
Get a paper towel and wipe it up
Wipe it up with your sock
GraphJam.com

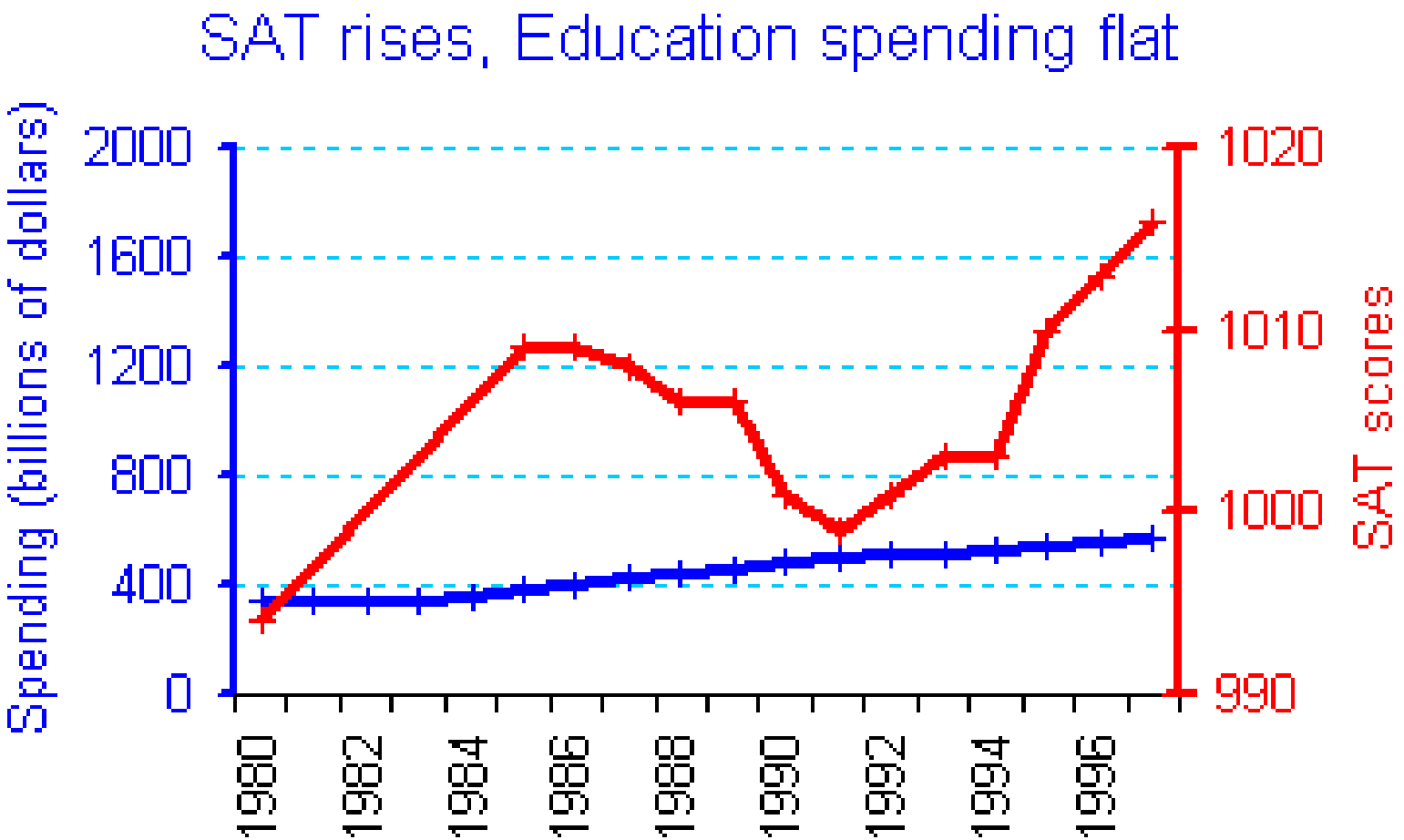
What's wrong...



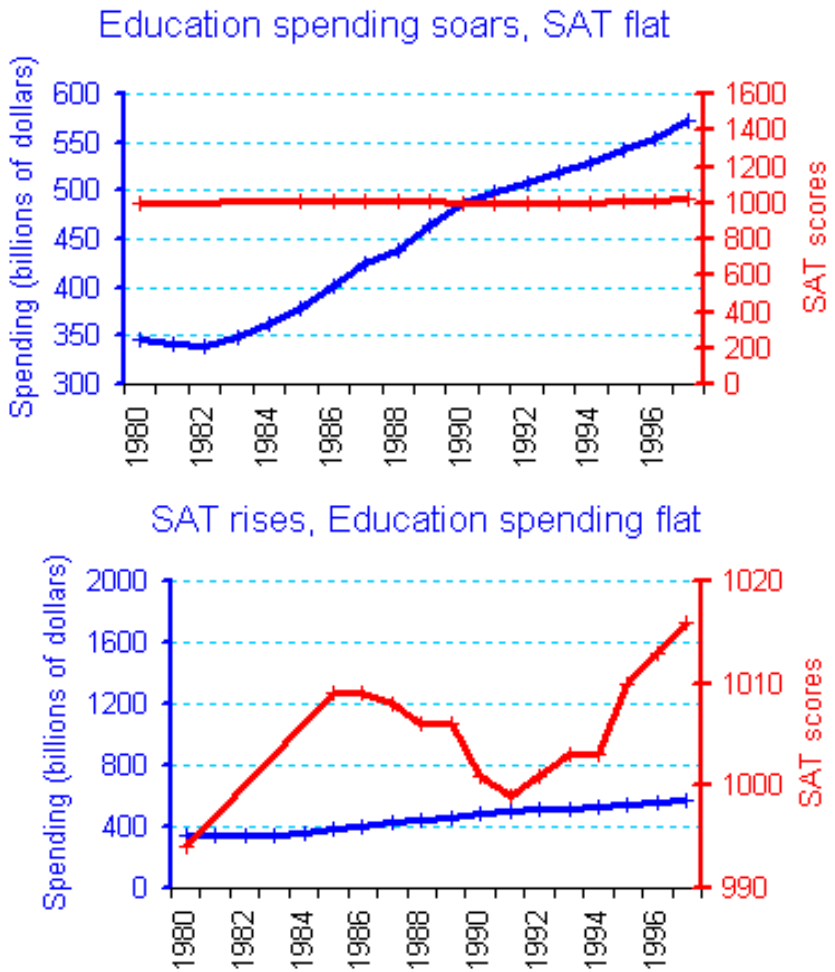
What's wrong...


Education spending soars, SAT flat
Spending (billions of dollars)
600
550
500
450
400
350
300
1600
1400
1200
1000
800
600
400
200
0
SAT scores
1980
1982
1984
1986
1988
1990
1992
1994
1996

What's wrong...


SAT rises, Education spending flat
Spending (billions of dollars)
2000
1600
1200
800
400
0
1020
1010
1000
990
SAT scores
1980
1982
1984
1986
1988
1990
1992
1994
1996

What's wrong...

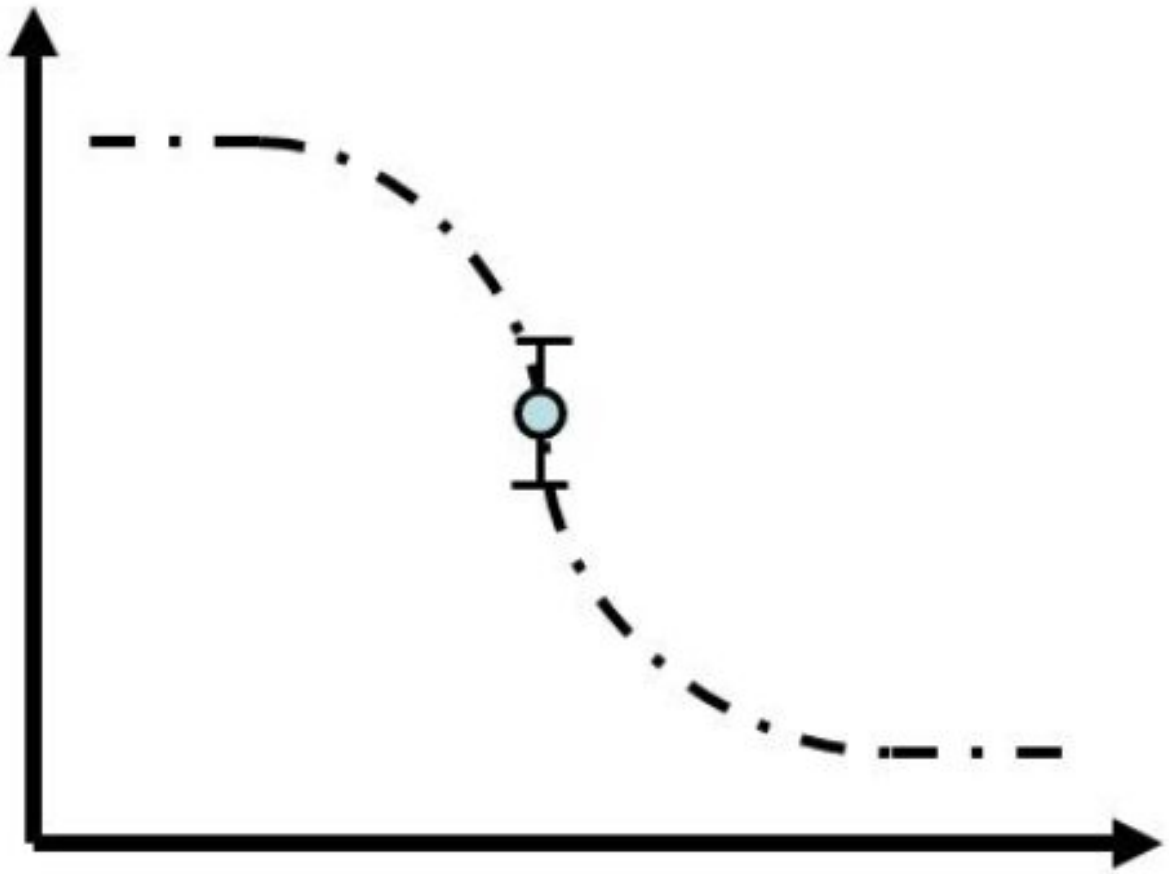

Education spending soars, SAT flat
Spending (billions of dollars)
600
550
500
450
400
350
300
1600
1400
1200
1000
800
600
400
200
0
SAT scores
1980
1982
1984
1986
1988
1990
1992
1994
1996
SAT rises, Education spending flat
Spending (billions of dollars)
2000
1600
1200
800
400
0
1020
1010
1000
990
SAT scores
1980
1982
1984
1986
1988
1990
1992
1994
1996

Bad Graphs

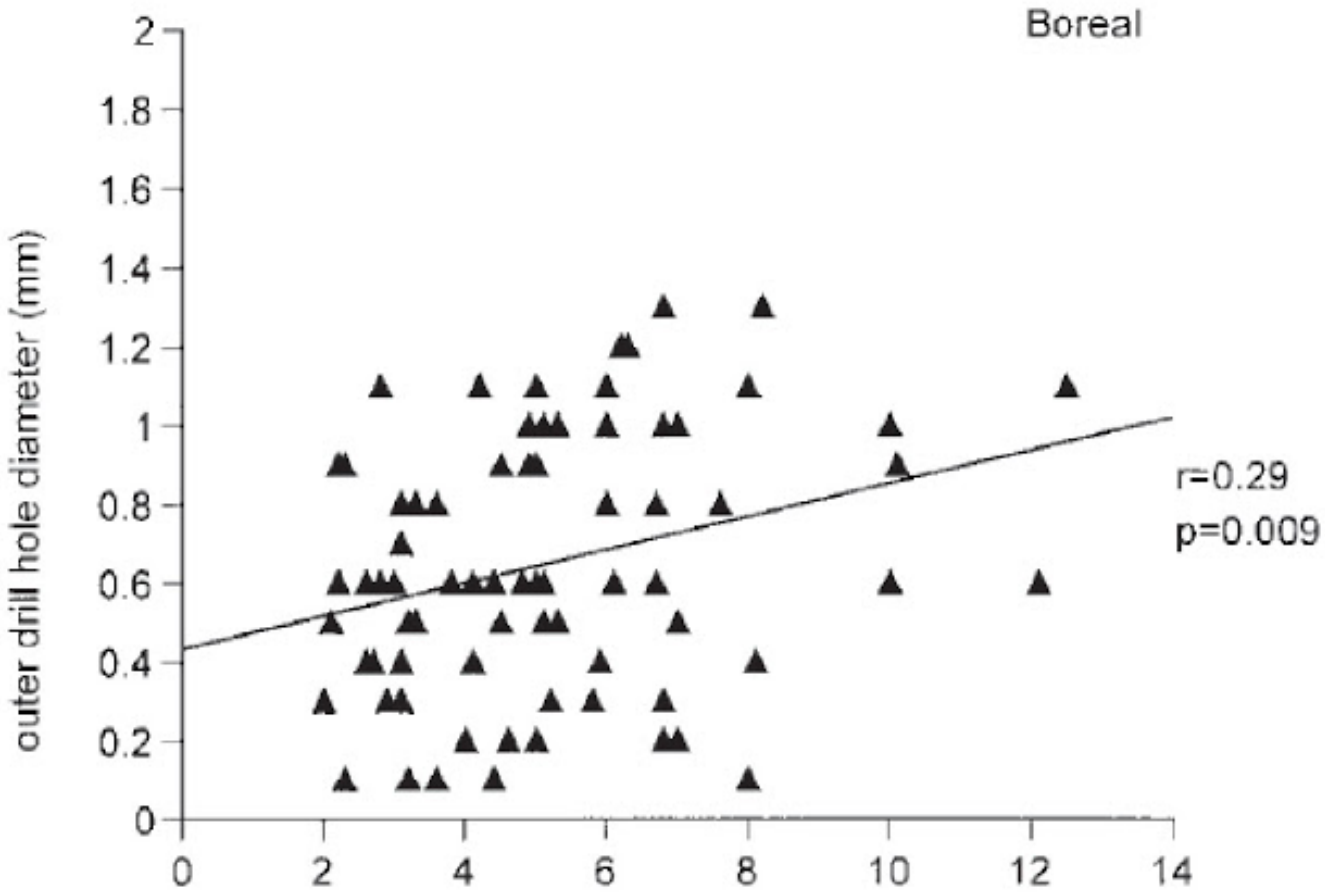
However, sometimes it's not so funny...

The next three graphs are from actual scientific articles and presentations at prestigious conferences and publications

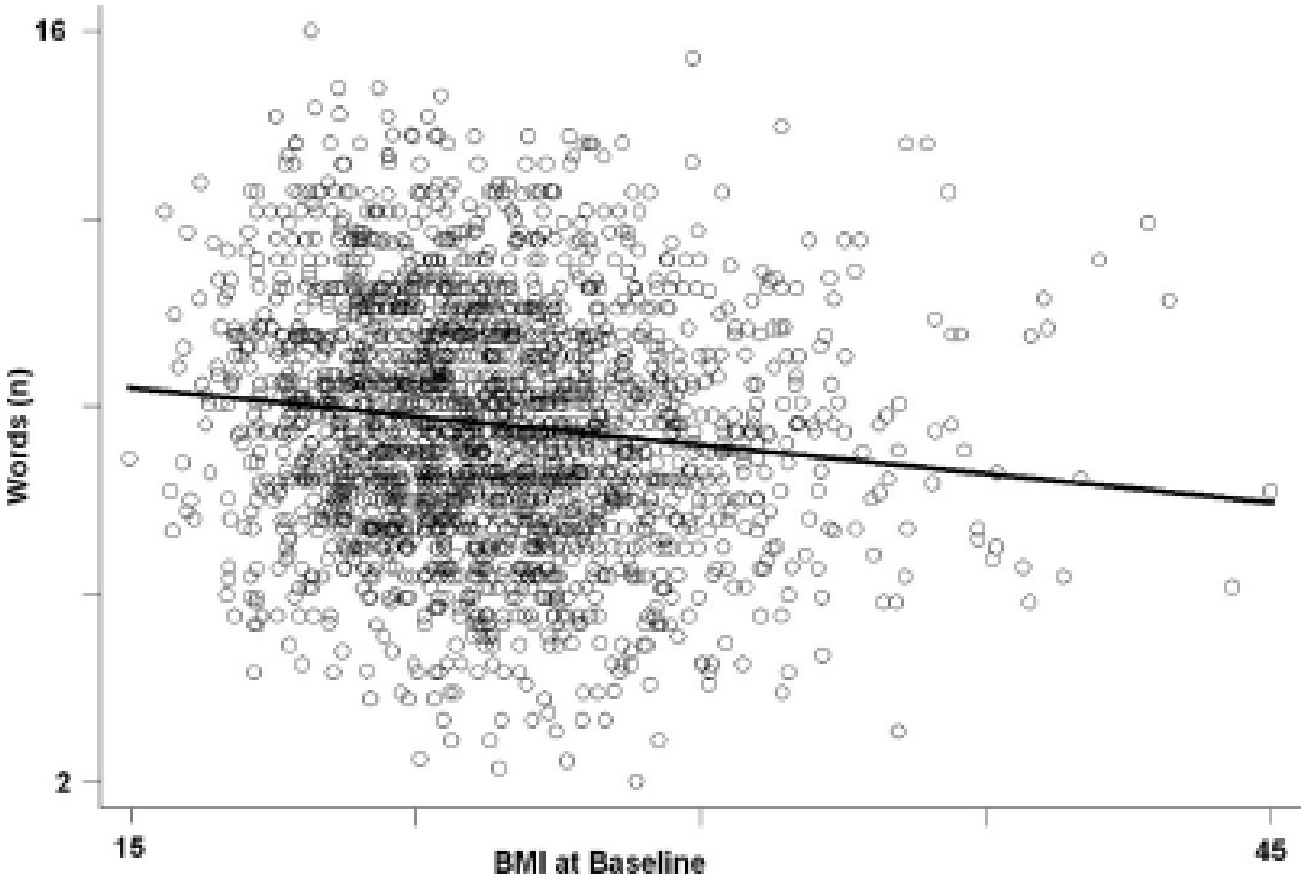
What's wrong...



What's wrong...


Boreal
outer drill hole diameter (mm)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
2
4
6
8
10
12
14
r=0.29
p=0.009

What's wrong...



Cleveland's Guidance

Clear Vision

- Make the data stand out. Avoid superfluity
- Do not clutter the data region
- Do not overdo the number of tick marks
- Use a reference line when there is an important value
- Do not allow data labels to interfere or clutter the graph
- Avoid putting notes, keys, and markers in the data region
- Superposed data sets must be readily visually discriminated
- Overlapping plotting symbols must be visually distinguishable
- Visual clarity must be preserved under reduction and reproduction

Cleveland's Guidance

Clear Understanding

- Error bars should be clearly explained
- When logarithms of a variable are graphed, the scale label should correspond to the tick mark labels
- Choose the range of the tick marks to include or nearly include the range of the data
- Subject to the constraints that scales have, choose the scales so that the data fill up as much of the data region as possible
- Do not insist that zero always be included on a scale showing magnitude [except for bar charts]
- Logarithmic scales can be your friend!

Gelman's Guidance

Graph issues

- Graphs are gimmicks
- Trade-off: the snazzier your display, the more you can get away with a crappy underlying analysis.
- Trust your peers enough to present their estimates and standard errors directly, with no tricks.
- Leave the dot plots, pie charts, moving zip charts, and all the rest to the marketing department and the art directors of Newsweek and USA Today.

Graph vs Table

Decision Questions [ncsu.edu/labwrite]

- Are the variables qualitative or quantitative?
- What is the total number of data points to be shown?
- Is there more than one independent variable?
- Are you trying to represent the statistical distribution of the data?
- How important is it to be able to see individual values?
- How important is it to understand the overall trend?

Which Graph?

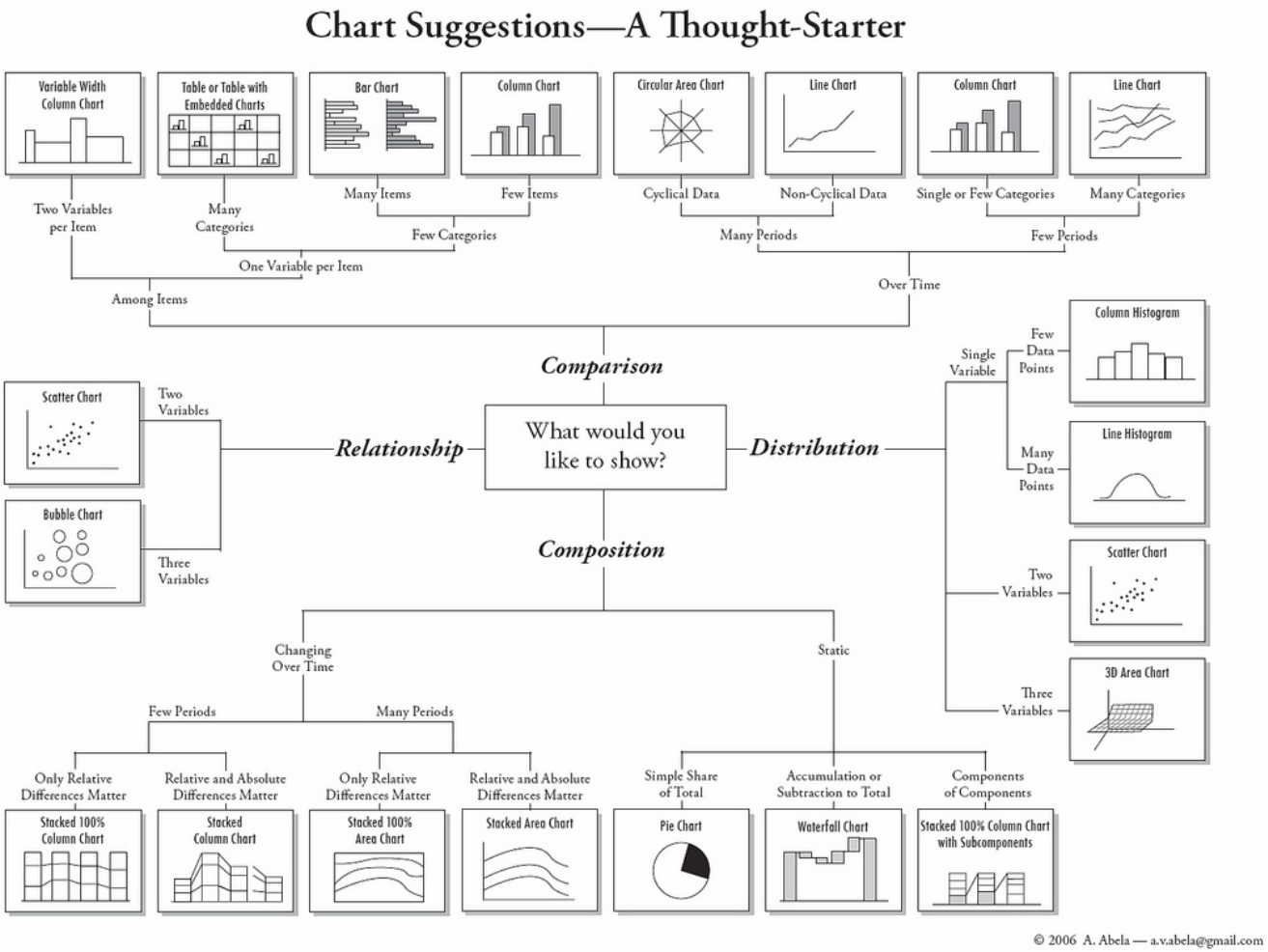


Chart Suggestions—A Thought-Starter
Variable Width Column Chart
Table or Table with Embedded Charts
Bar Chart
Column Chart
Circular Area Chart
Line Chart
Column Chart
Line Chart
Two Variables per Item
Many Categories
Many Items
Few Items
Cyclical Data
Non-Cyclical Data
Single or Few Categories
Many Categories
Few Categories
Many Periods
Few Periods
One Variable per Item
Over Time
Among Items
Comparison
Scatter Chart
Two Variables
Relationship
What would you like to show?
Distribution
Single Variable
Few Data Points
Column Histogram
Many Data Points
Line Histogram
Bubble Chart
Three Variables
Composition
Two Variables
Scatter Chart
Changing Over Time
Static
Three Variables
3D Area Chart
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Stacked 100% Column Chart
Stacked Column Chart
Stacked 100% Area Chart
Stacked Area Chart
Pie Chart
Waterfall Chart
Stacked 100% Column Chart with Subcomponents
© 2006 A. Abela — a.v.abela@gmail.com

Basic R Graphical Parameters

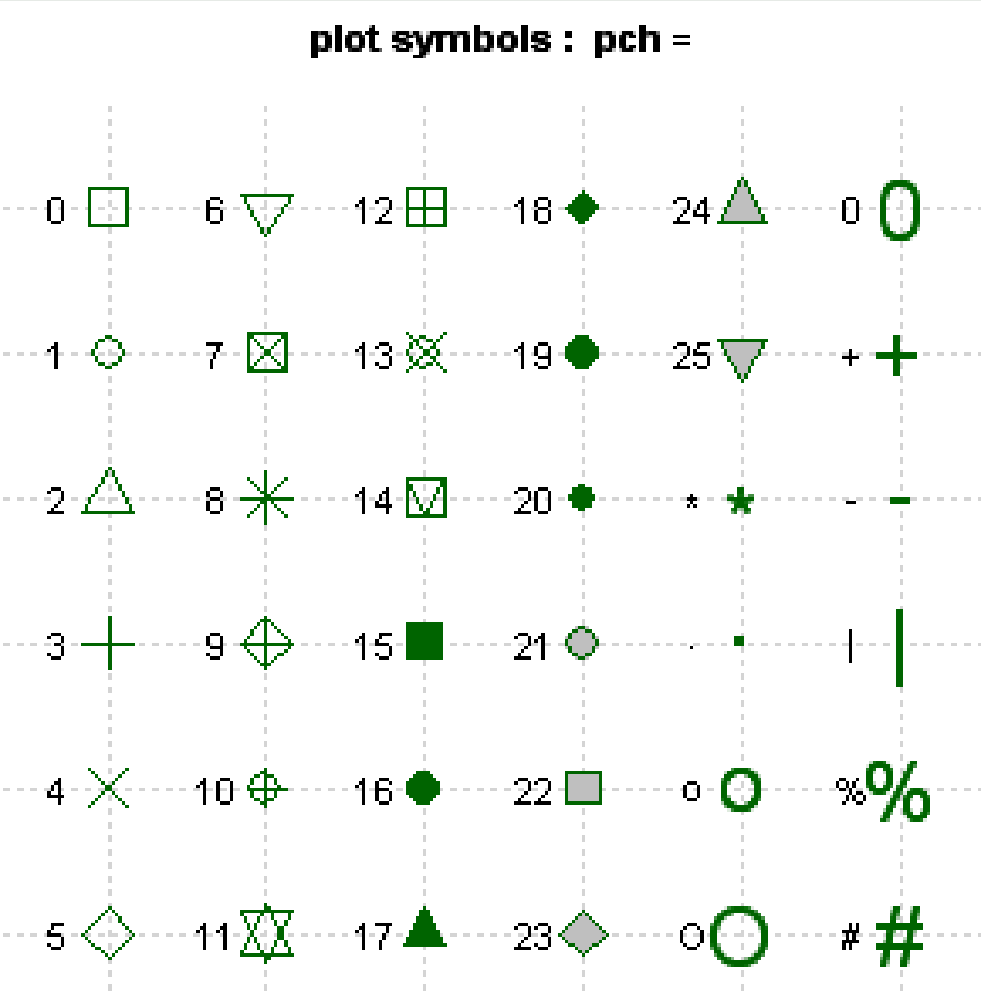
Big Idea

R graphical items, like **plot** and **hist** has the ability to add options. You can either change them globally with a **par** command, or you can do it individually on each graphical item

Let's look at some of the more common ones

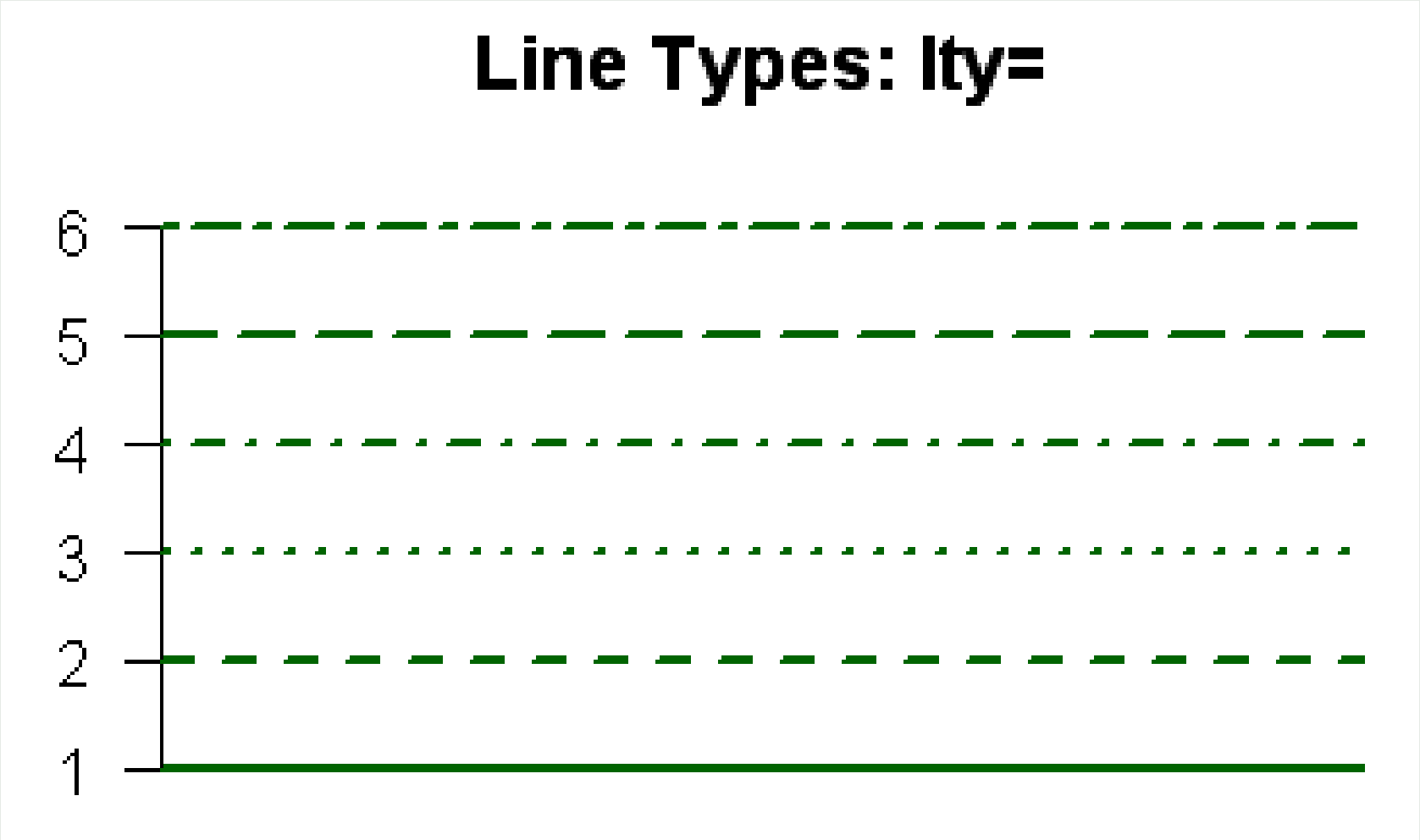
Basic R Graphical Parameters

Plotting symbols, pch=



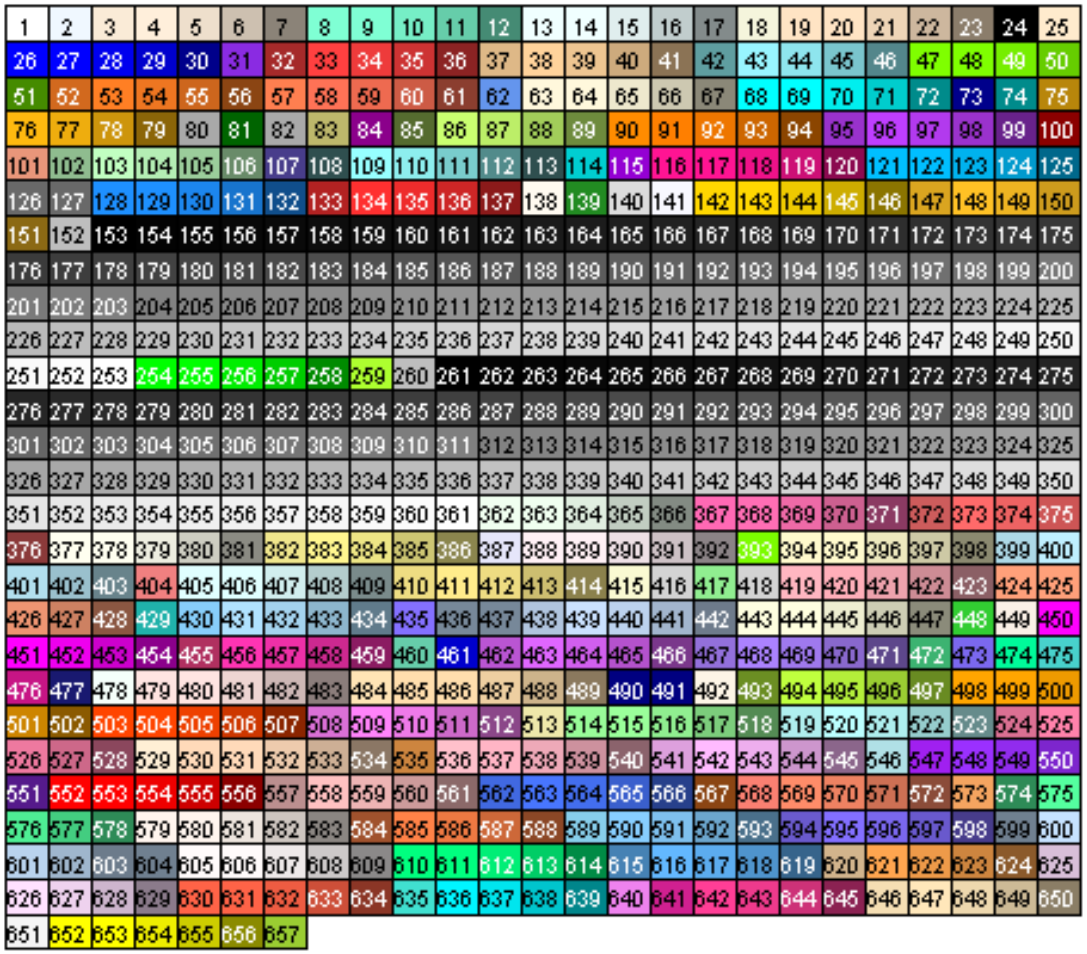
Basic R Graphical Parameters

Lines, lty= and lwd=



Basic R Graphical Parameters

Colors, col=, bg=, and fg=

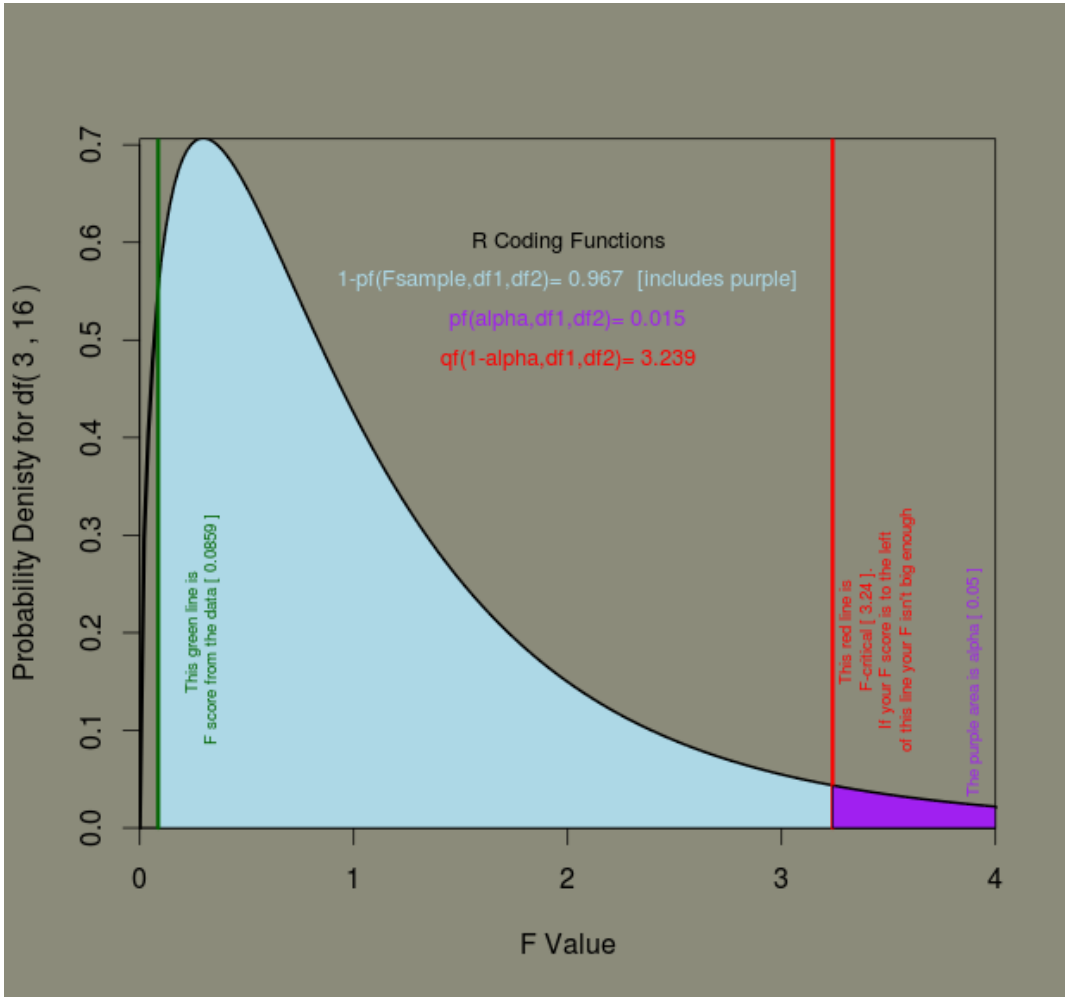


Basic R Graphical Parameters

Axes, Titles, and Legends

```
1     # Specify axis options within plot()
2 plot(x, y, main="title", sub="subtitle",
3   xlab="X-axis label", ylab="y-axix label",
4   xlim=c(xmin, xmax), ylim=c(ymin, ymax))
5
6   title(main="main title", sub="sub-title",
7    xlab="x-axis label", ylab="y-axis label")
8
9    legend(location, title, legend, ...)
```

F Score Example

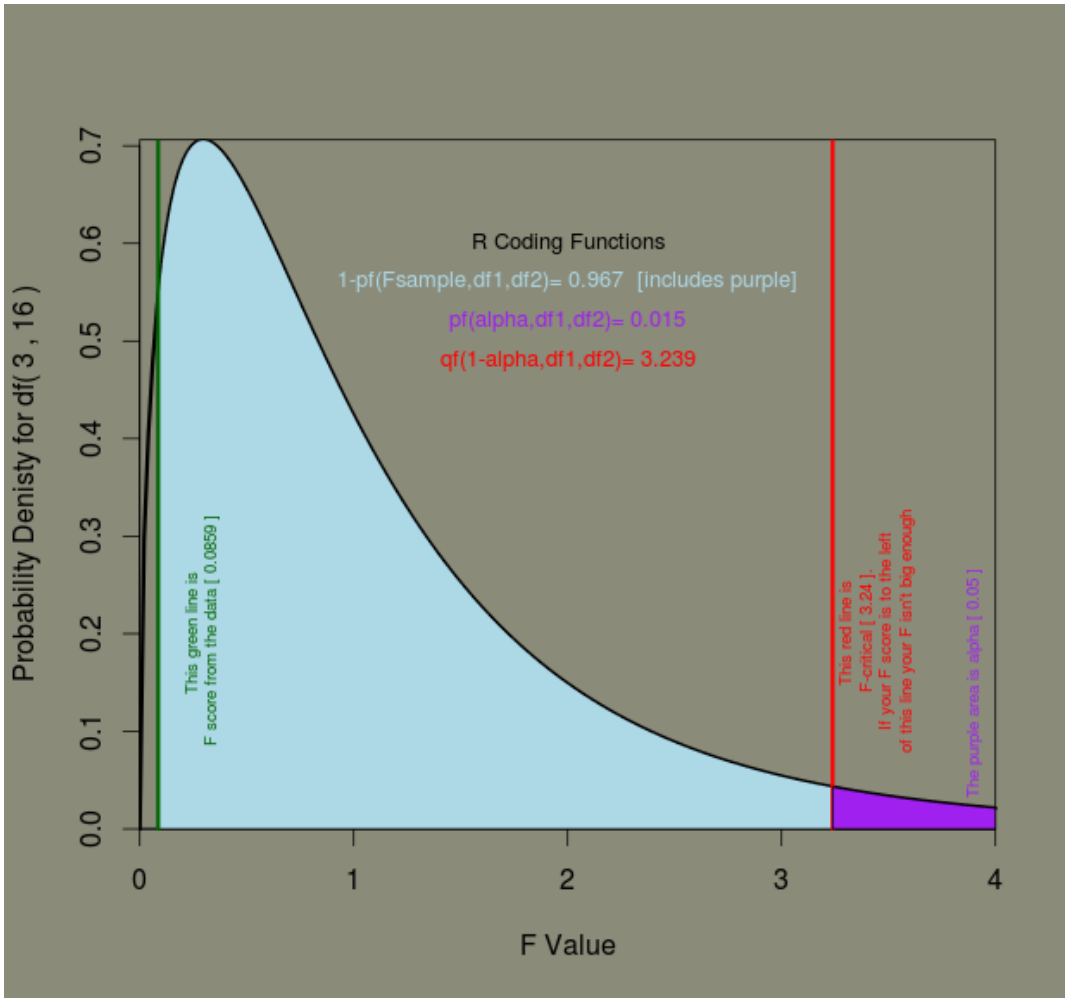


F Score Example

```
1    maxFplot <-4 # The maximum you want the F curve ploted
2    alpha <- 0.05 #The significance level you want
3    df1 <- 3 #degrees of freedom for the numerator
4    df2 <- 16 #degrees of freedom for the denominator
5    Fsample <- 0.0859
```

Some preliminaries for the context

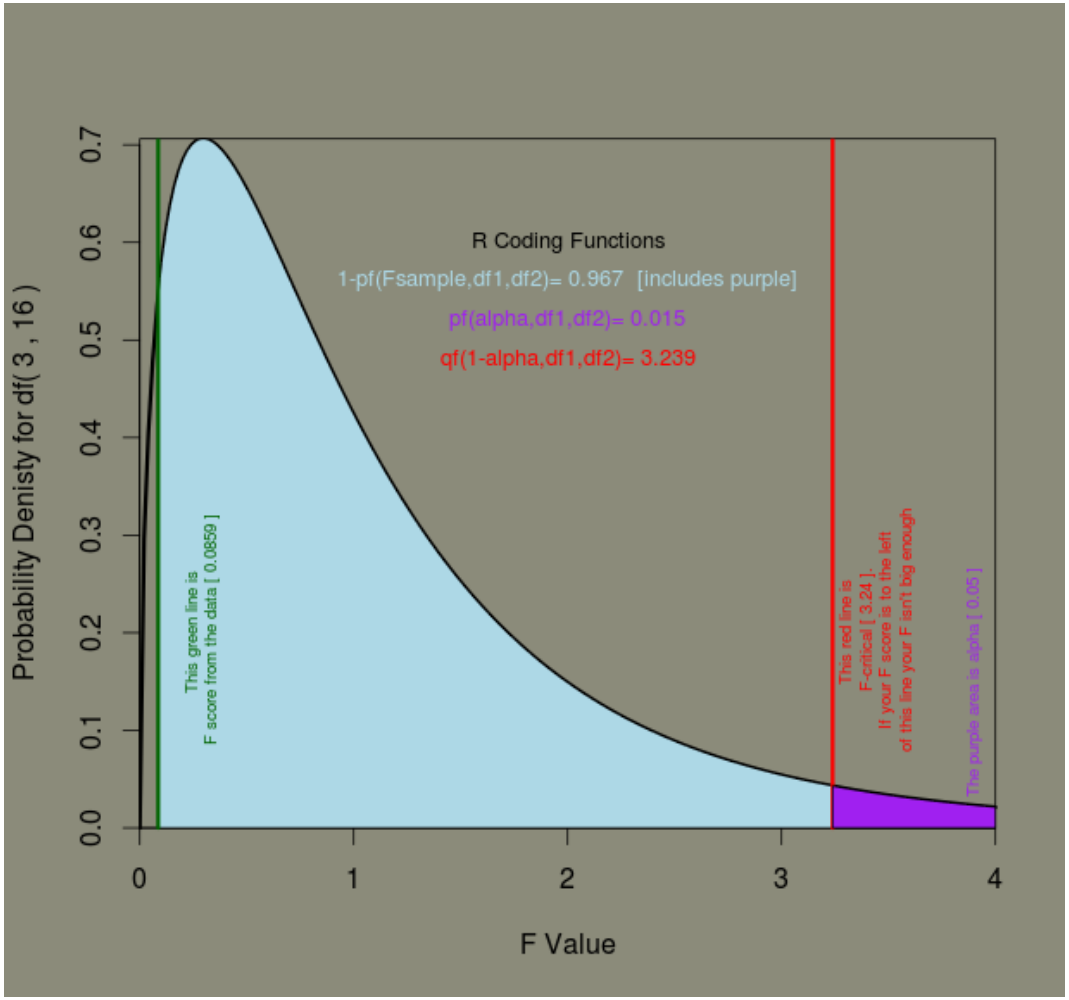
F Score Example



F Score Example - Setup and F Curve Plot

```
1  par(bg = "lightyellow4")
2  x=seq(0,maxFplot,length=200)
3  y=df(x,df1,df2)
4  plot.new()
5  plot(yaxs="i",xaxs="i", x,y,type="l",lwd=3,col="black",ylab=paste("
       Probability Denisty for df(",df1,",",df2,")"),xlab="F Value")
```

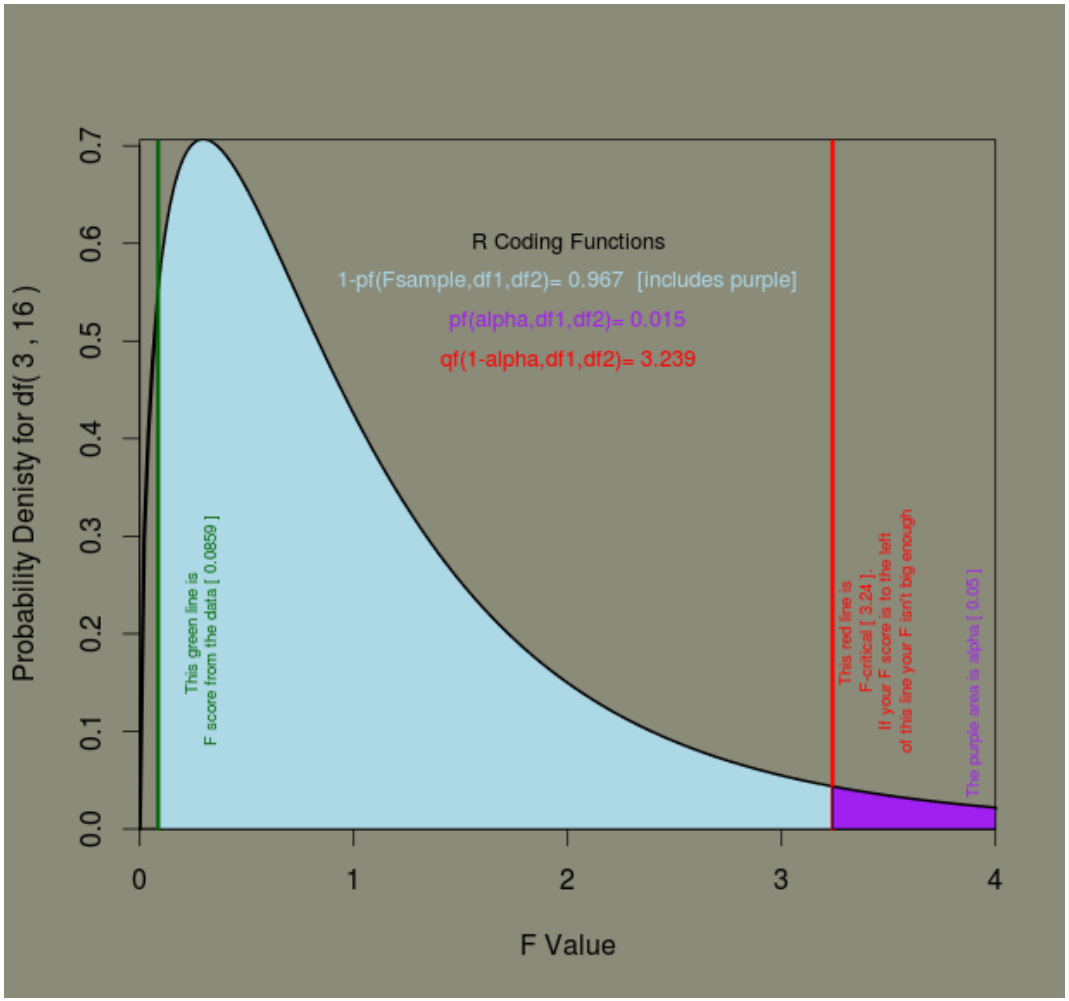
F Score Example



F Score Example - Coloring in the area between

```
1  x=seq(Fsample, maxFplot,length=200)
2  y=df(x,df1,df2)
3  polygon(c(Fsample, x, maxFplot),c(0,y,0),col="lightblue")
```

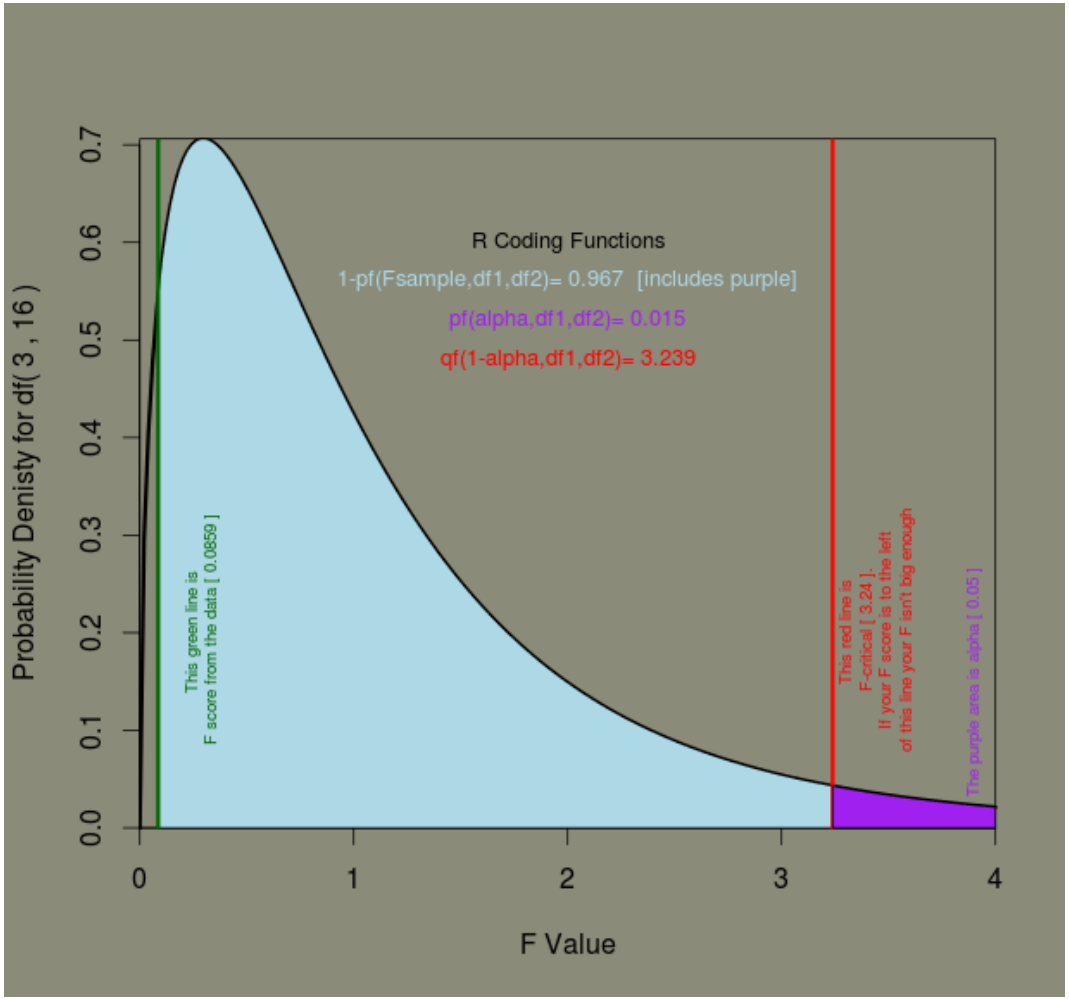
F Score Example



F Score Example - Red and green reference lines

```
1  abline(v=qf(1-alpha,df1,df2), lwd=3, col="red")
2  abline(v=Fsample, lwd=3, col="darkgreen")
```

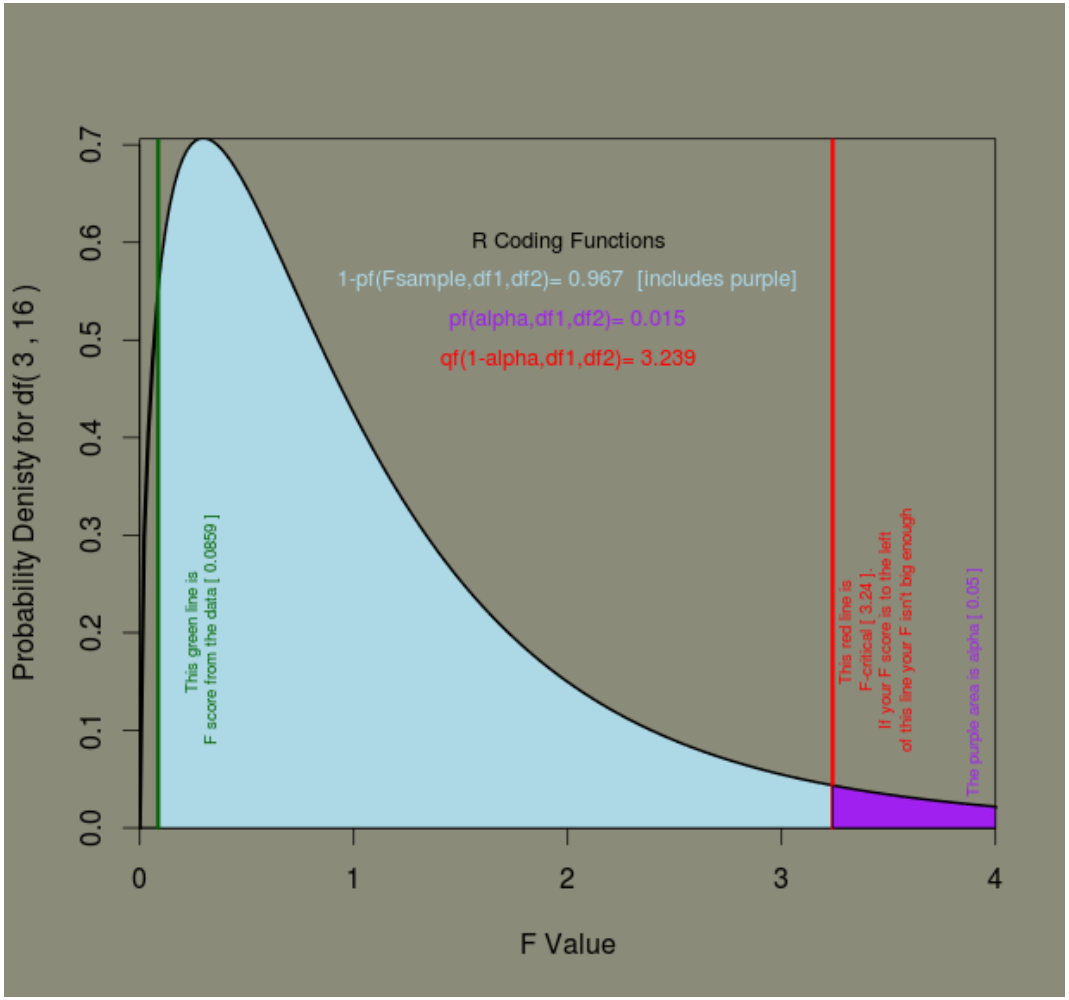
F Score Example



F Score Example -The purple alpha region

```
x=seq(qf(1-alpha,df1,df2),maxFplot,length=200)
y=df(x,df1,df2)
polygon(c(qf(1-alpha,df1,df2),x,maxFplot),c(0,y,0),col="purple")
```

F Score Example



F Score Example -The text annotations

```
1 text(Fsample+0.2, 0.2,srt=90, labels=paste("This green line is\n F
      score from the data [",Fsample,"]"),col="darkgreen", cex=0.6)
2 text(maxFplot-0.1,0.15,srt=90,labels=paste("The purple area is alpha ["
      ,alpha,"]"),col="purple", cex=0.6)
3 text(qf(1-alpha,df1,df2)+0.2, srt=90, 0.2,labels=paste("This red line
      is\n F-critical [",round(qf(1-alpha,df1,df2),2),"].\n If your F
      score is to the left\n of this line your F isn't big enough"),col="
      red", cex=0.6)
4 text(maxFplot/2, 0.6,labels="R Coding Functions",col="black", cex=0.8)
5 text(maxFplot/2, 0.56,labels=paste("1-pf(Fsample,df1,df2)=",round(1-pf(
      Fsample,df1,df2),3)," [includes purple]"),col="lightblue", cex=0.8)
6 text(maxFplot/2, 0.52,labels=paste("pf(alpha,df1,df2)=",round(pf(alpha,
      df1,df2),3)),col="purple", cex=0.8)
7 text(maxFplot/2, 0.48,labels=paste("qf(1-alpha,df1,df2)=",round(qf(1-
      alpha,df1,df2),3)),col="red", cex=0.8)
```

F Score Example