Plotting and Data Visualization for Biostatistical Consulting

Gulce Askin

August 1, 2017

OUTLINE

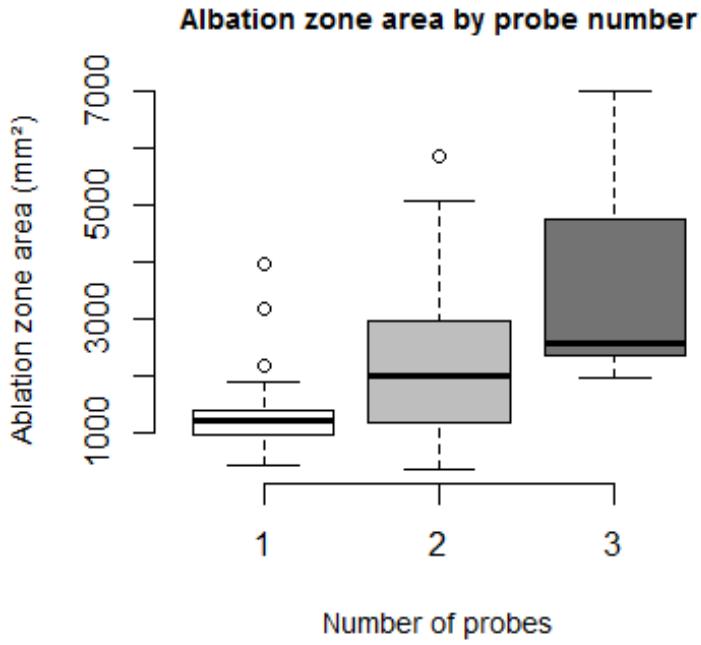
- I. Elements of clear, accurate, effective plots
- II. Useful graphics for biostatistical consultations & collaborations
- III. Resources

I. Elements of clear, accurate, effective plots

- Precise axis labels with units
  - i.e. “Blood glucose” vs “Blood glucose concentration in mg/dl”
  - Omit 0’s where possible
- Concise, informative titles
  - Should fully explain the graphic
  - Sub-title can be used to mention relevant units or time frames
    - i.e. “5-year progression-free survival”
- Brief captions for additional relevant data
  - Always include total and subgroup sample sizes
  - Details about method used

Avoid causal terminology

- We are generally working with associations
- Avoid active language in plot titles and labels
  - i.e. “The use of multiple probes increased the area of ablation”



Histogram bins

- Can have large impact on histogram appearance
- Goal is to display essential structure/distribution of the data
- R and SAS have default sizes
  - In base R hist() use ***breaks =***
  - In SAS proc univariate use ***midpoints=*** or ***endpoints=***

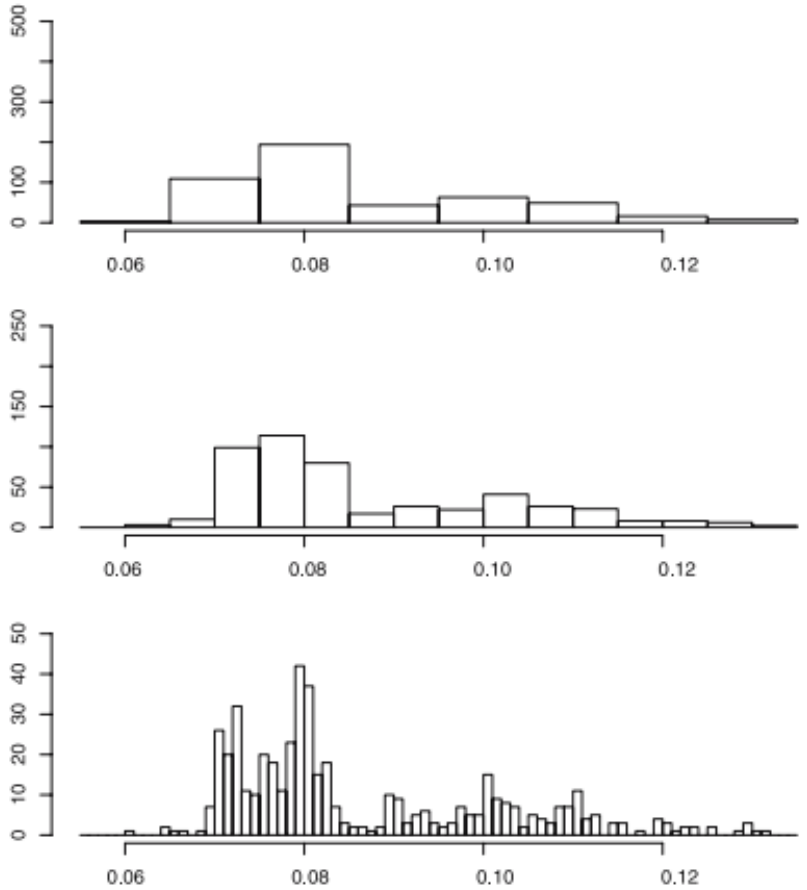
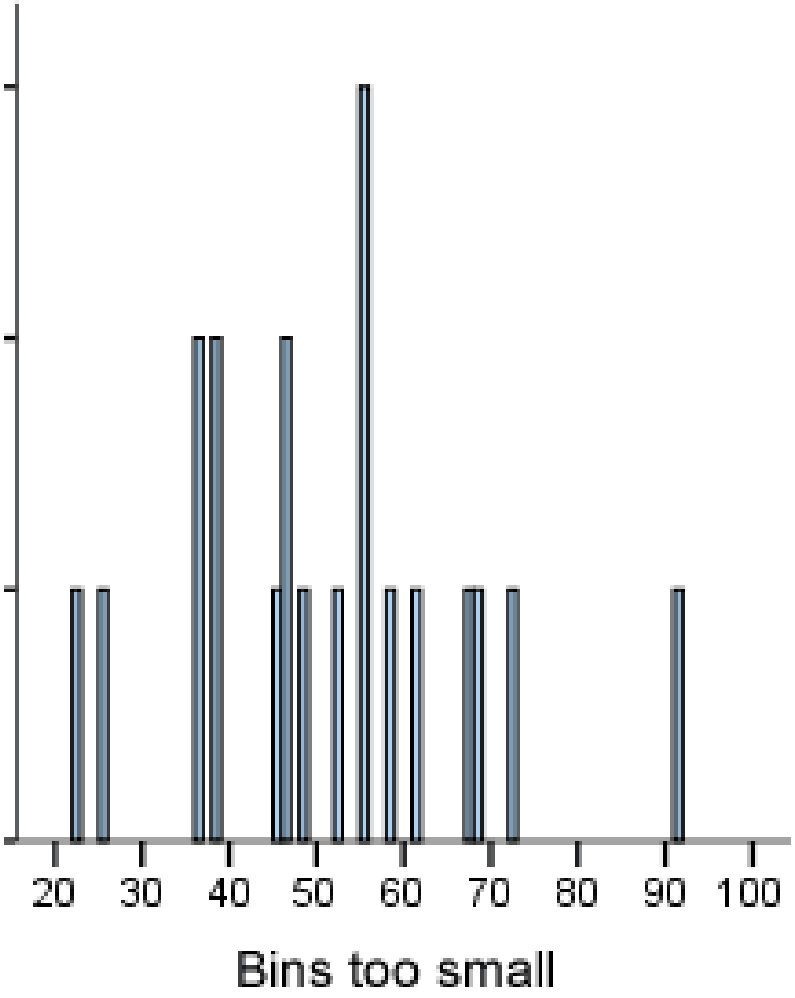
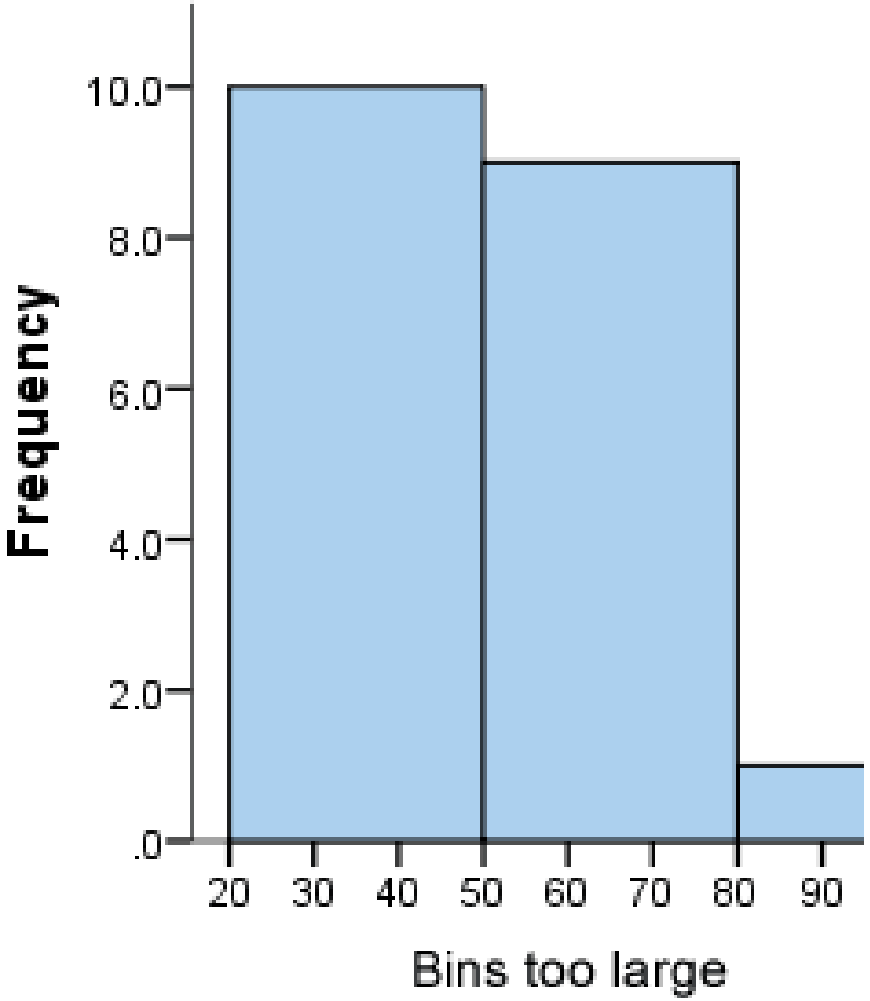


Figure 2.4. Three different histograms of the Hidalgo stamp thickness data, all with the same anchorpoint but with different binwidths. The *horizontal scales* are aligned and the total area of each display is the same (note the different frequency scales). Source: Izenman and Sommer (1988)

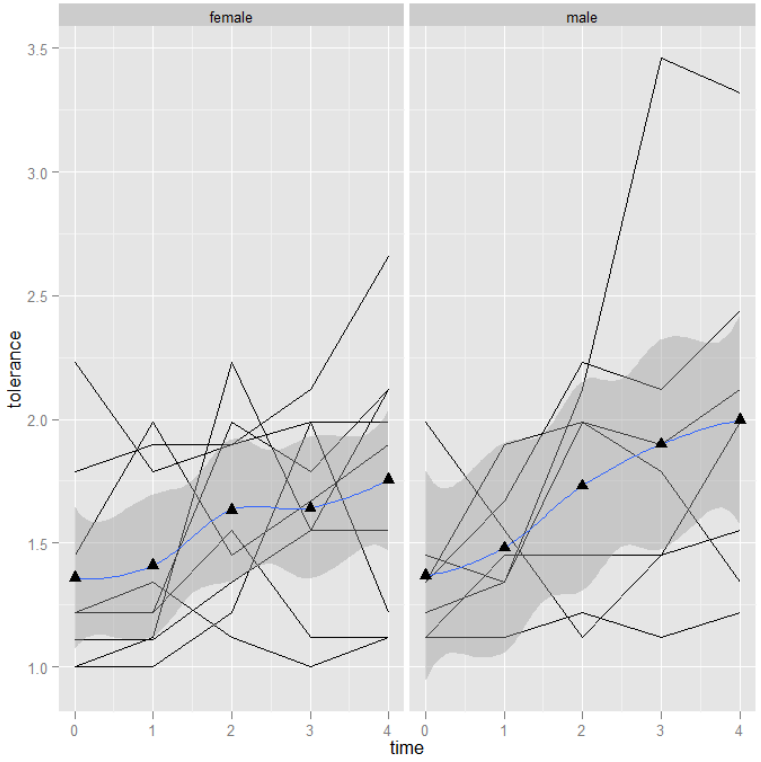
Histogram bins

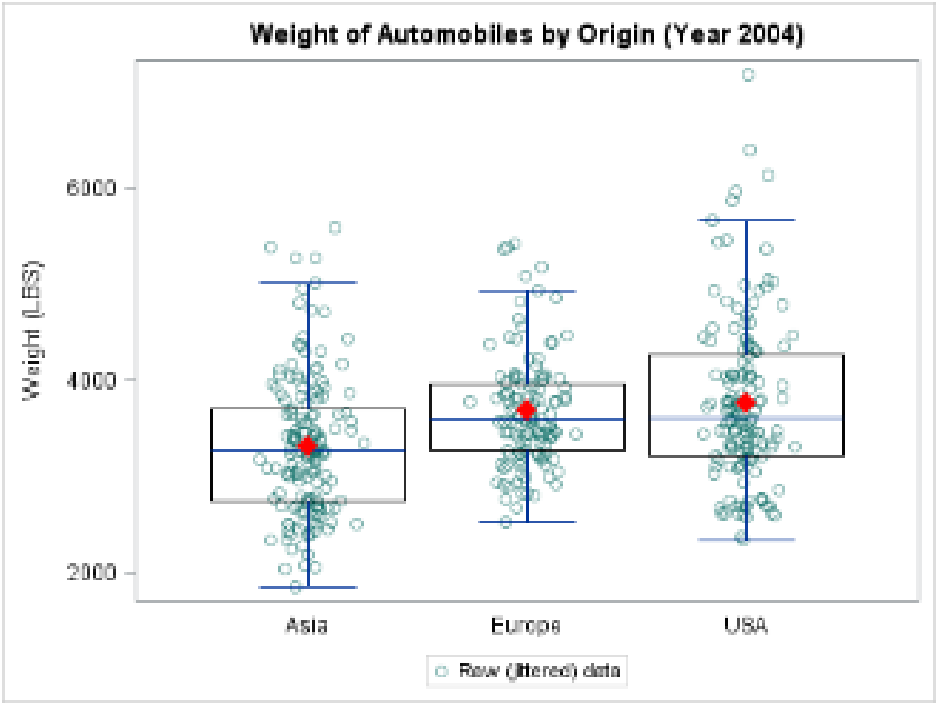




Show the data points

- Can work especially well for smaller samples
- Overlays provide information about the variability in the sample

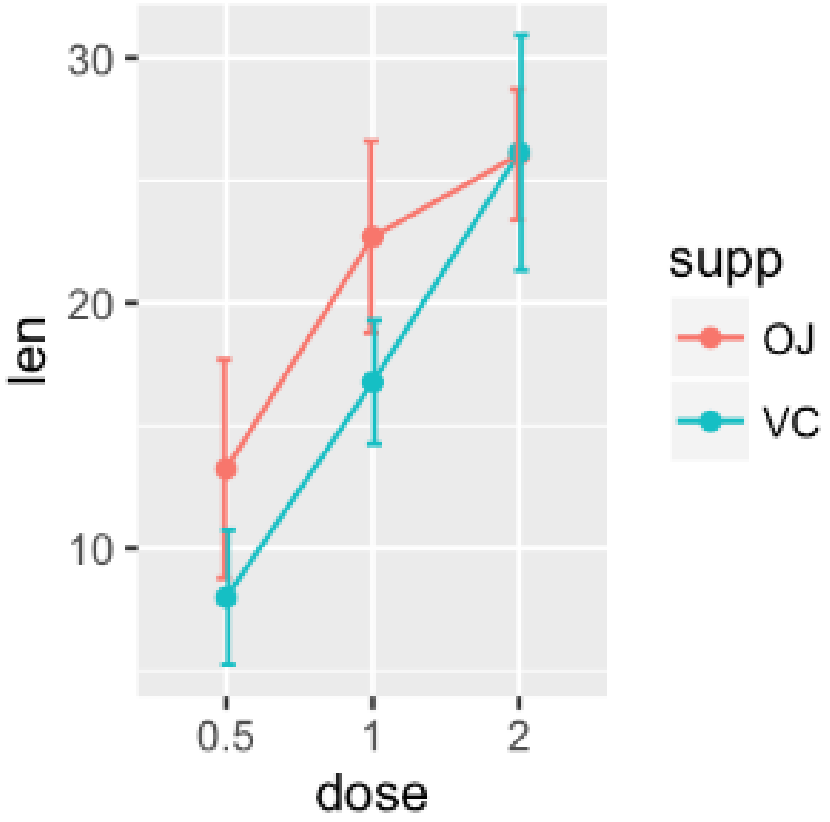




Show error/uncertainty

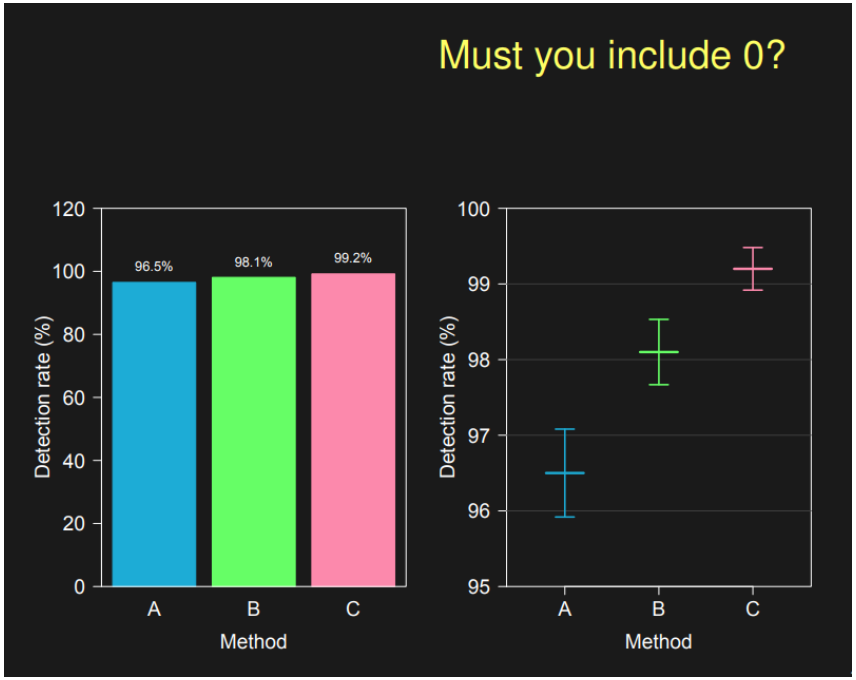
Three types of error bars

- Standard deviation
- Standard error
- Confidence interval (CI)
  - Most intuitive for readers
- If you use error bars, specify



Scale

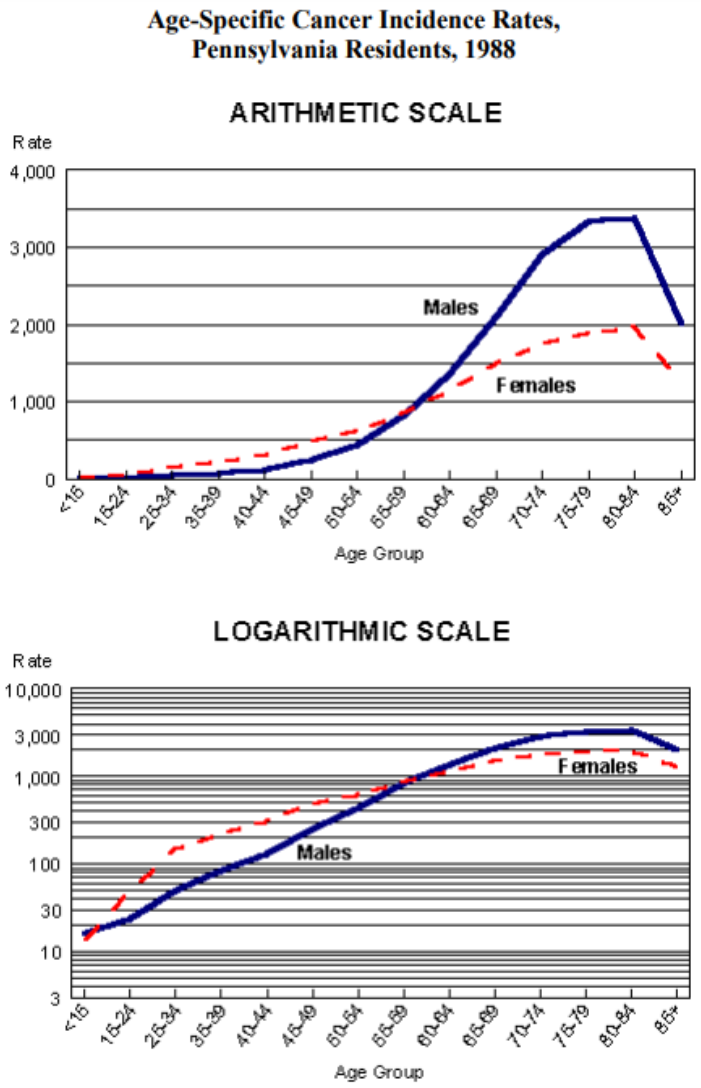
- Zero controversy
  - Does distribution contain negative values
  - Is zero a meaningful reference value?
  - Does graphing without zero artificially emphasize small differences among datapoints?
  - Does graphing at zero visually minimize meaningful differences among datapoints?



Scale

Logarithmic Scale

- Address skewness towards large values
- Useful when it is important to understand percent change or multiplicative factors
- Know your audience. Not all audiences are comfortable with logarithms
- The most common base is ten

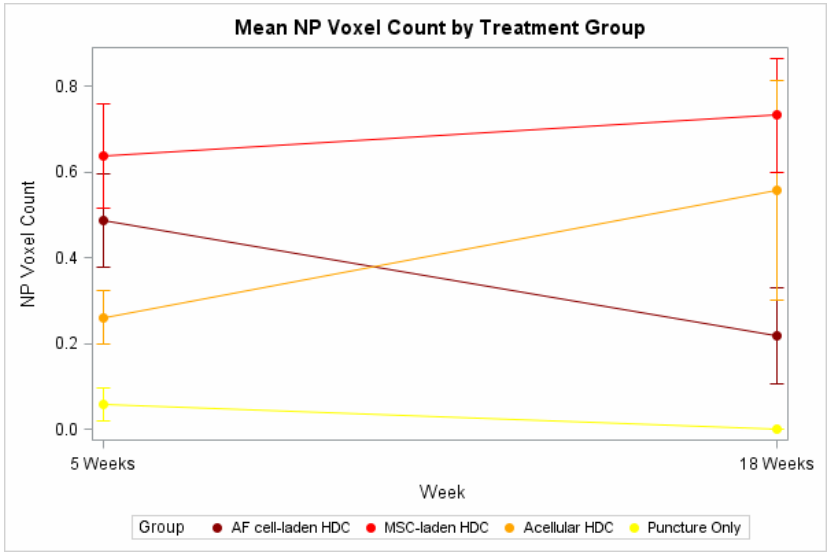


Other elements…

- Standalone
  - Do not make reader refer back to the text
  - Make sure each and every feature is identified
- Color
  - Do not distract from data
  - Try to select colorblind-friendly range, if possible
    - See informative resource at end
  - Include markers for curves for curves for greyscale printing

Color – a bit of fun

- R package
- “wesanderson”



Give your R charts that Wes Anderson style

I'm a big fan of Wes Anderson's movies. I love the quirky characters and stories, the distinctive cinematography, and the unique visual style. Now you can bring some of that style to your own R charts, by making use of these Wes Anderson inspired palettes. Just choose your favourite Wes Anderson film or short:

Install the wesanderson pallettes package, created by Karthik Ram:

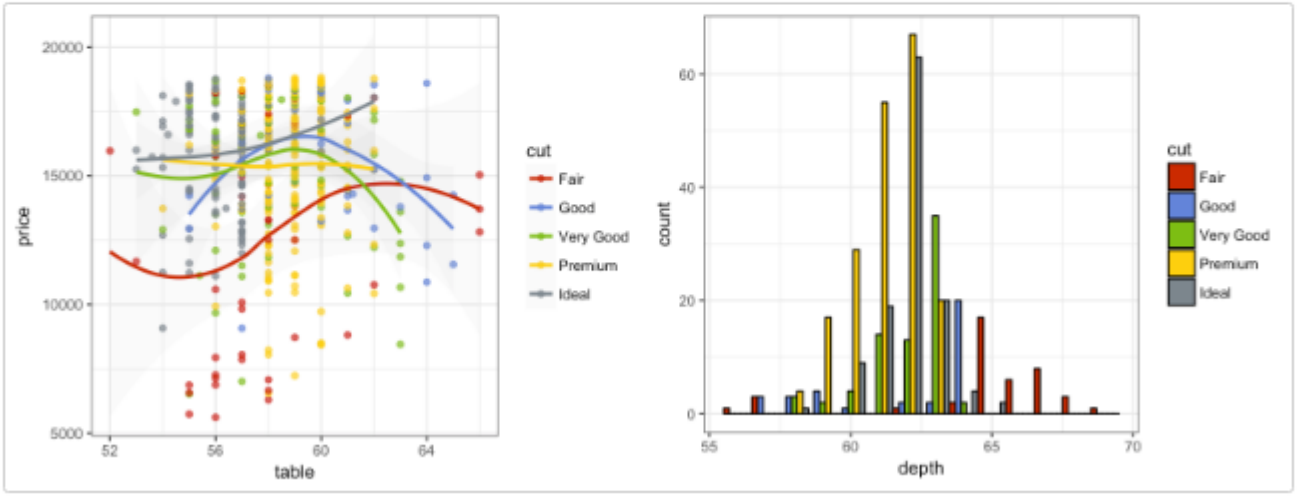
Color – a bit of fun

- Scientific Journal and Sci-Fi Themed Color Palettes for ggplot2

2.12 Star Trek

This palette is inspired by the (uniform) colors in *Star Trek*:

```
p1_startrek = p1 + scale_color_startrek()
p2_startrek = p2 + scale_fill_startrek()
grid.arrange(p1_startrek, p2_startrek, ncol = 2)
```



Metrics for bad data displays

Tufte (1983)

- Lie factor

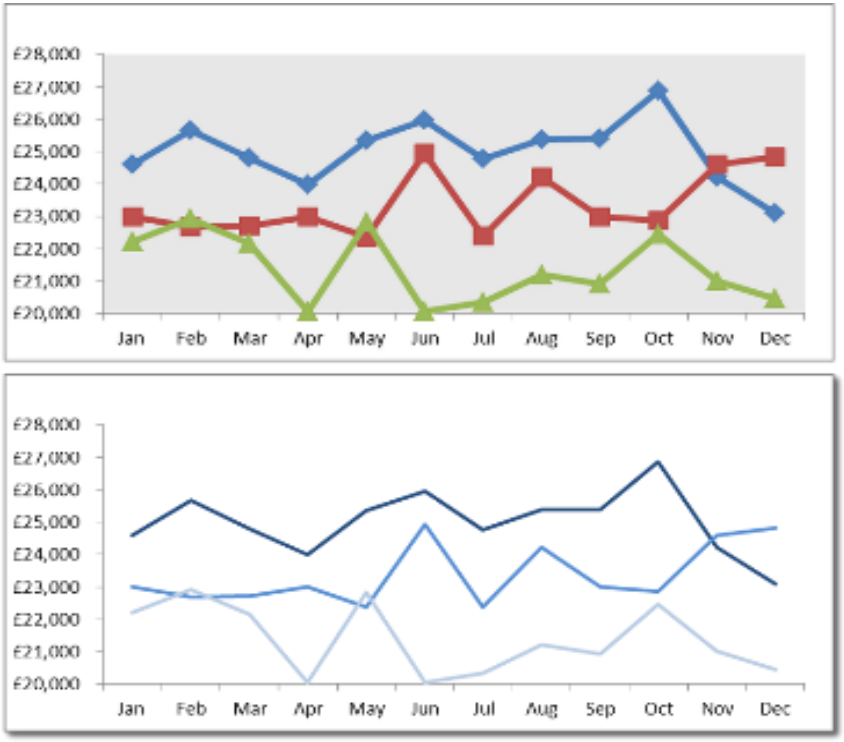
$$lie\ factor = \frac{size\ of\ effect\ shown\ in\ graphic}{size\ of\ effect\ in\ data}$$

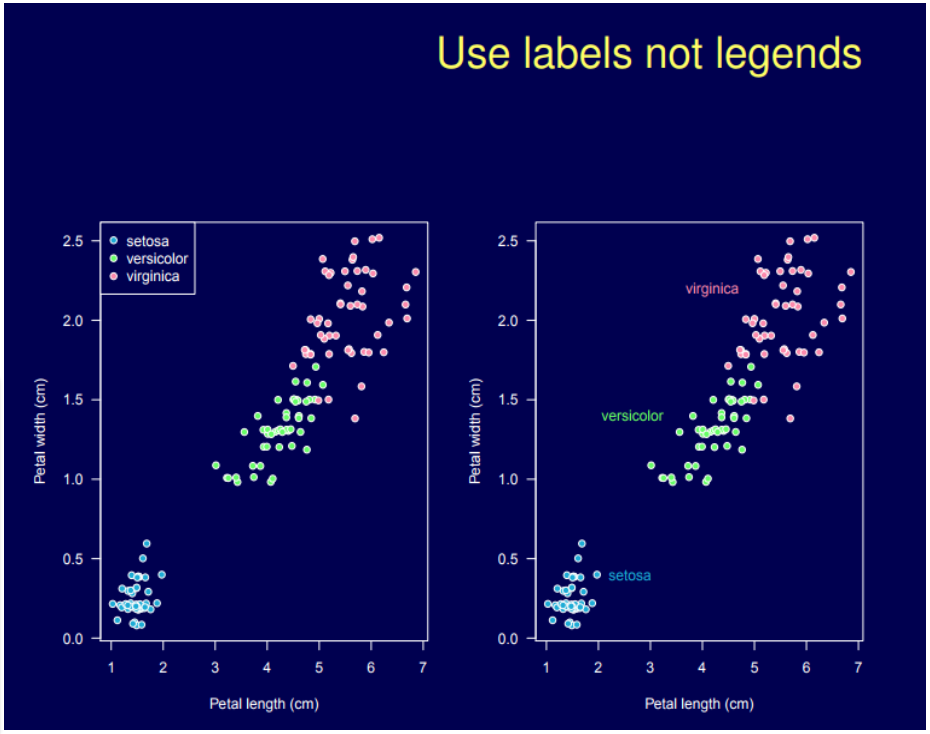
- Data density index

$$ddi = \frac{\#\ entries\ in\ data\ matrix}{area\ of\ data\ graphic}$$

Use your judgement

- Legends or data labels?
- Border or no border?

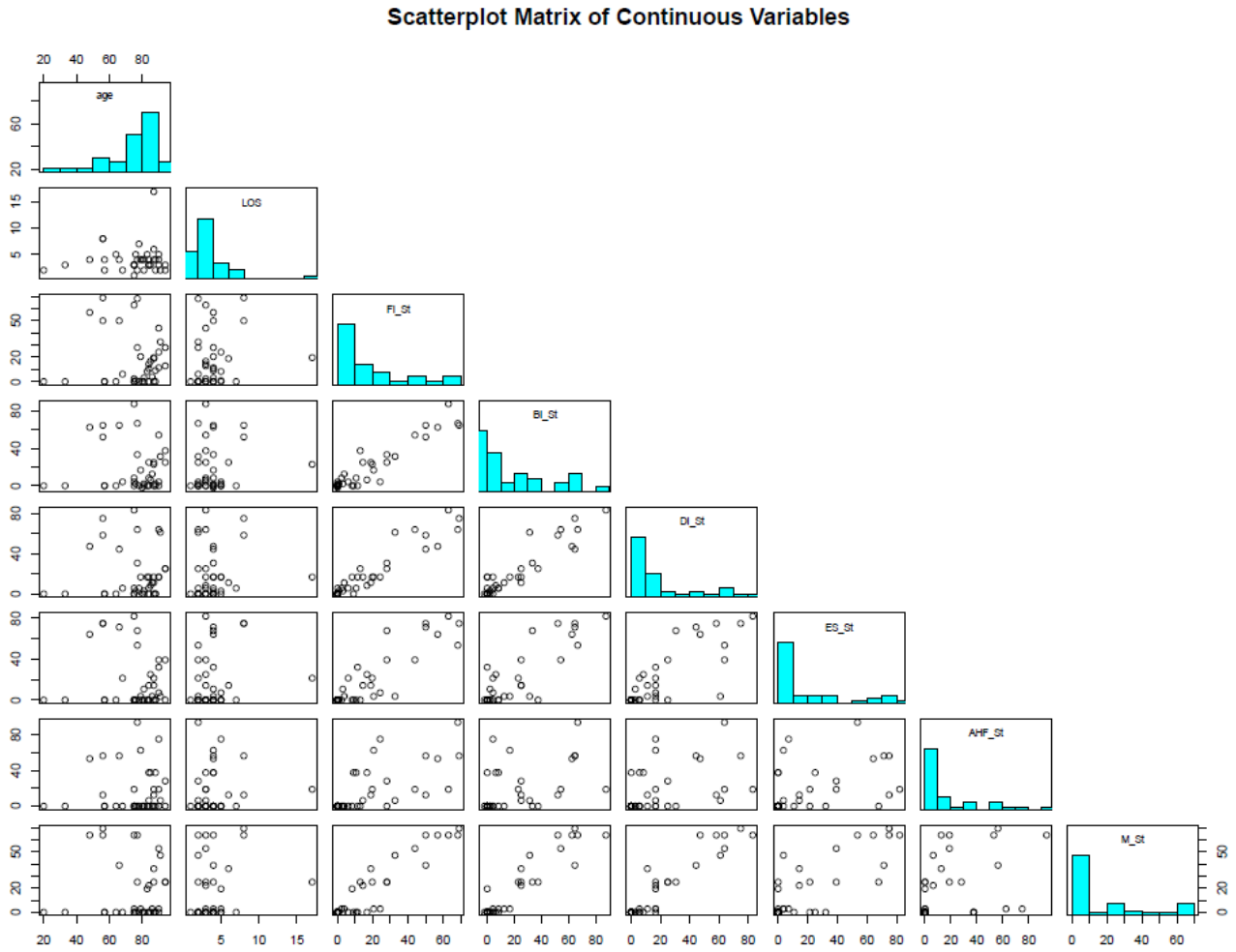




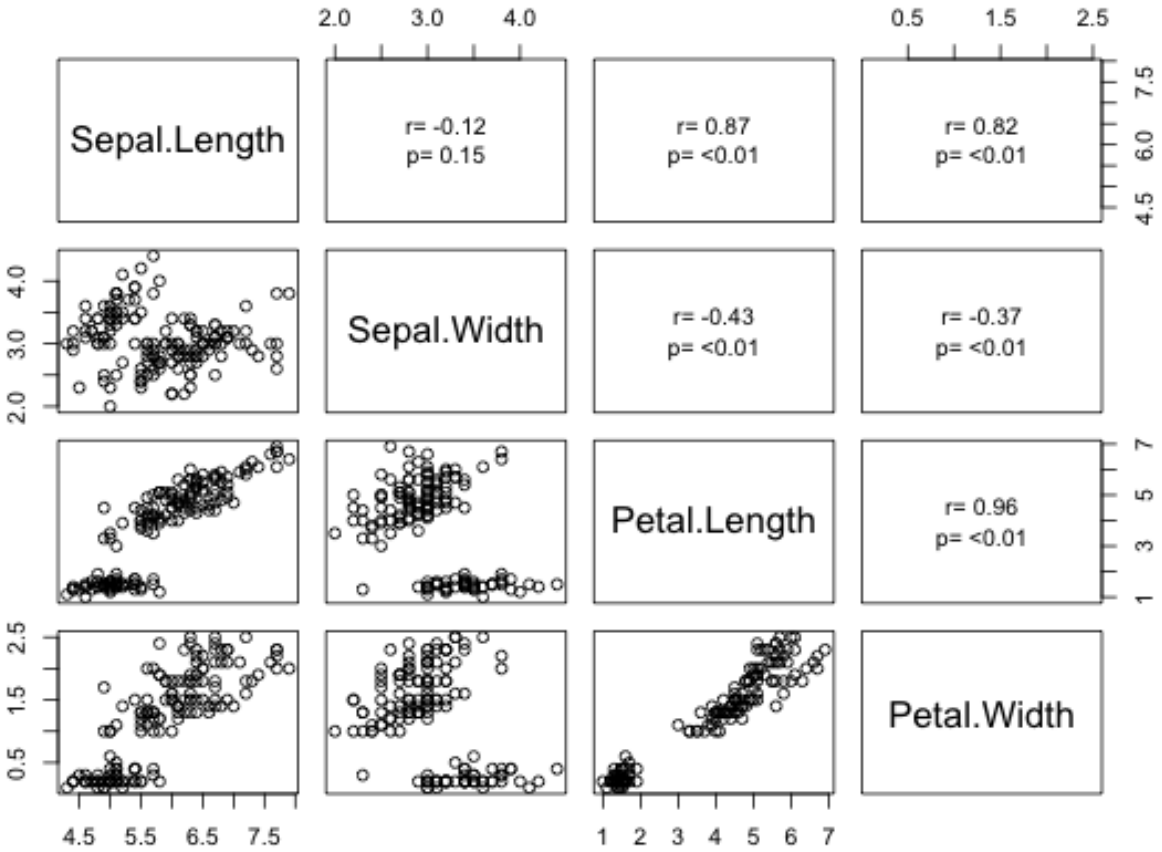
II. Useful graphics for biostatistical consultations & collaborations

- Plots will vary by purpose
  - For your report: communicate results to collaborator
  - For publication: communicate to wider audience

Scatterplot matrix



Scatterplot matrix



Scatterplot matrix

```
panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
}

age_corr_panel<-
pairs(~age+LOS+FI_St+BI_St+DI_St+ES_St+AHF_S
t+M_St,data=lco, main="Scatterplot Matrix of
Continuous Variables", diag.panel=panel.hist,
upper.panel= NULL)
```

```
panel.cor <- function(x, y, digits = 2, cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  # correlation coefficient
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste("r= ", txt, sep = "")
  text(0.5, 0.6, txt)

  # p-value calculation
  p <- cor.test(x, y)$p.value
  txt2 <- format(c(p, 0.123456789), digits = digits)[1]
  txt2 <- paste("p= ", txt2, sep = "")
  if(p<0.01) txt2 <- paste("p= ", "<0.01", sep = "")
  text(0.5, 0.4, txt2)
}
pairs(iris, upper.panel = panel.cor)
```

Scatterplot matrix

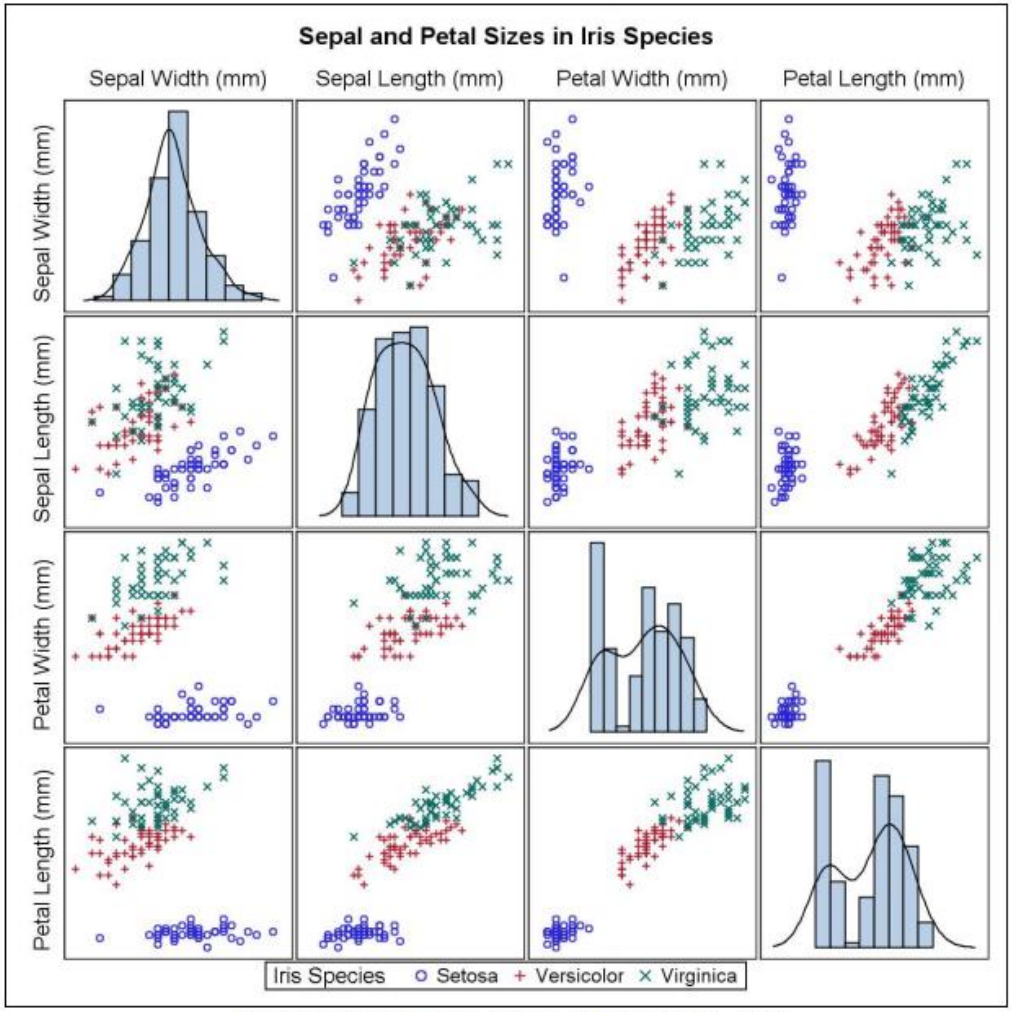


Figure 19: The MATRIX statement using the DIAGONAL option

```
title1 "Sepal and Petal Sizes in Iris Species";

proc sgscatter data=sashelp.iris;
        matrix sepalwidth
        sepallength petalwidth
        petallength / group=species
diagonal=(histogram kernel);

run;
```

Mosaic plot

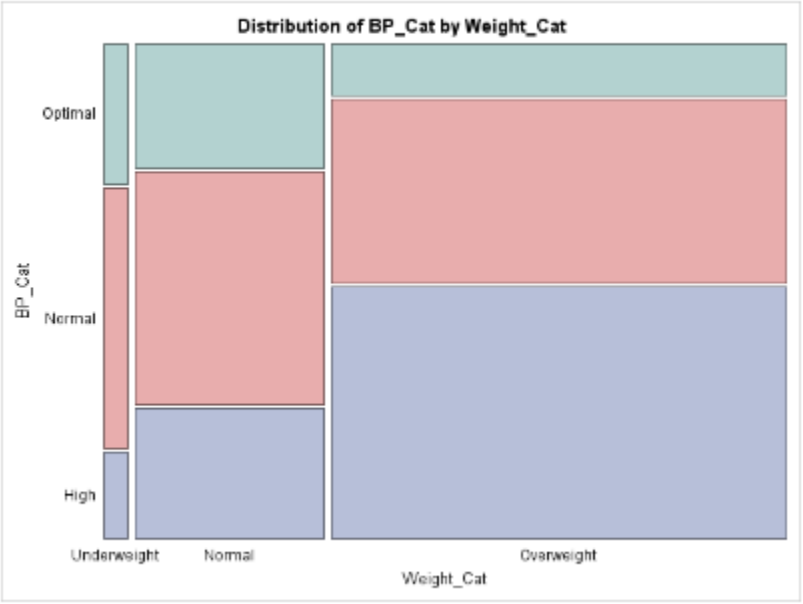
Display frequencies

Frequency
Percent
Col Pct

Table of BP_Cat by Weight_Cat

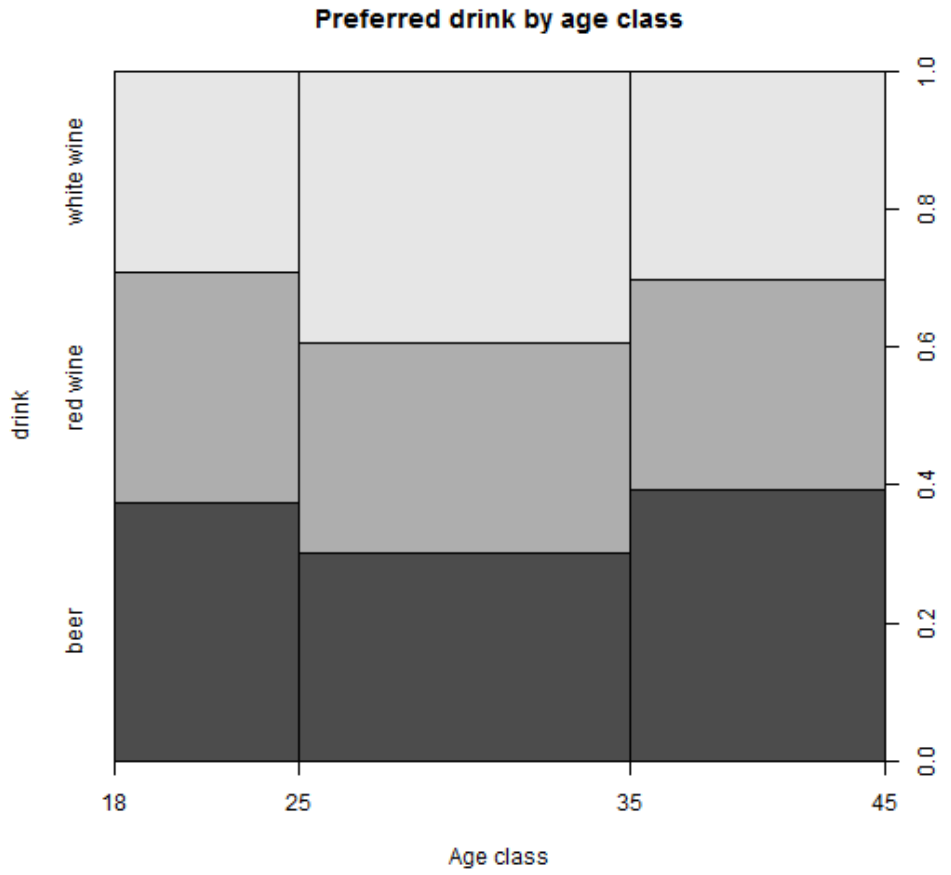
BP_Cat	Weight_Cat			
	Underweight	Normal	Overweight	Total
Optimal	52 1.00 28.73	374 7.19 25.41	371 7.13 10.45	797 15.32
Normal	97 1.86 53.59	704 13.53 47.83	1340 25.75 37.75	2141 41.15
High	32 0.62 17.68	394 7.57 26.77	1839 35.34 51.80	2265 43.53
Total	181 3.48	1472 28.29	3550 68.23	5203 100.00

Frequency Missing = 6



Mosaic plot

ageCls <- cut(age, breaks=lims, labels=LETTERS[1:(length(lims)-1)])
group <- factor(sample (letters[1:2], N, replace=TRUE))
cTab <- table(ageCls, pref, group) mosaicplot(cTab, cex.axis=1)



Graphical display for frequencies

Raters' characterization of sentences (absolute score range)	*Percent*
Negative (1–1.9)	23.9
Neutral (2–2.9)	52.2
Positive (≥ 3)	23.9
Total	100.0

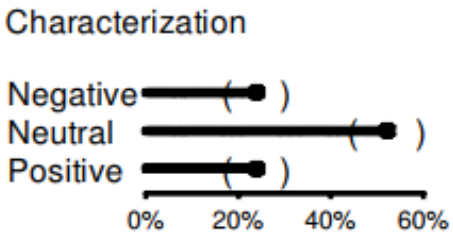


Figure 1. Top panel: Table from Ellenberg (2000) shows the relative frequencies of three categories in a set of 46 ratings of sentences. Bottom panel: Graphical display allows direct comparison without distractions of irrelevant decimal places. Parentheses show ±1 standard error bounds based on the implicit binomial distribution with n = 46.

Pie charts

- help("pie")

Note

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

Cleveland (1985), page 264: “Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements.” This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.

Patient trajectories

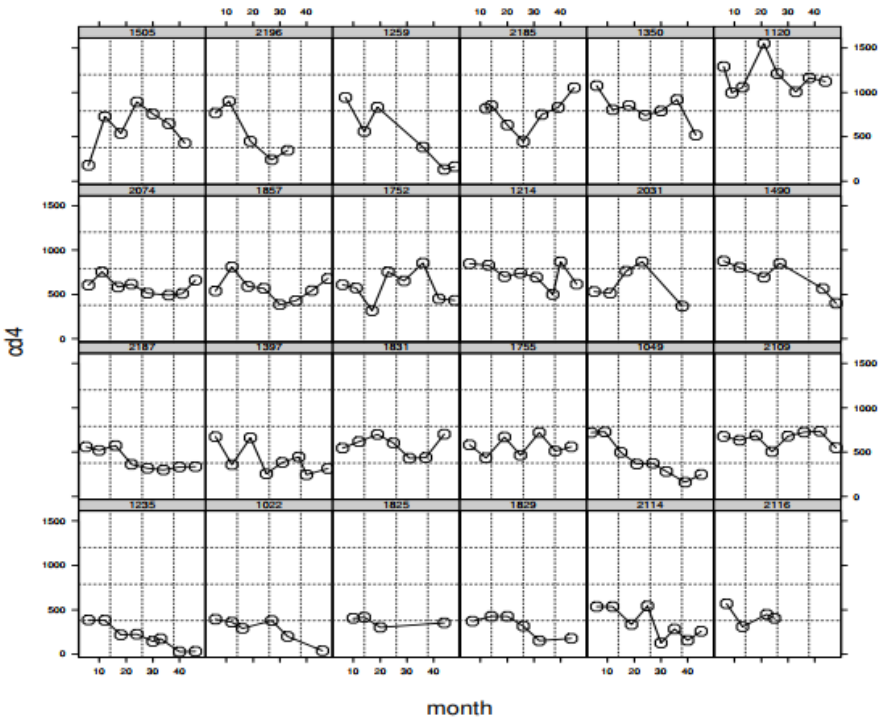
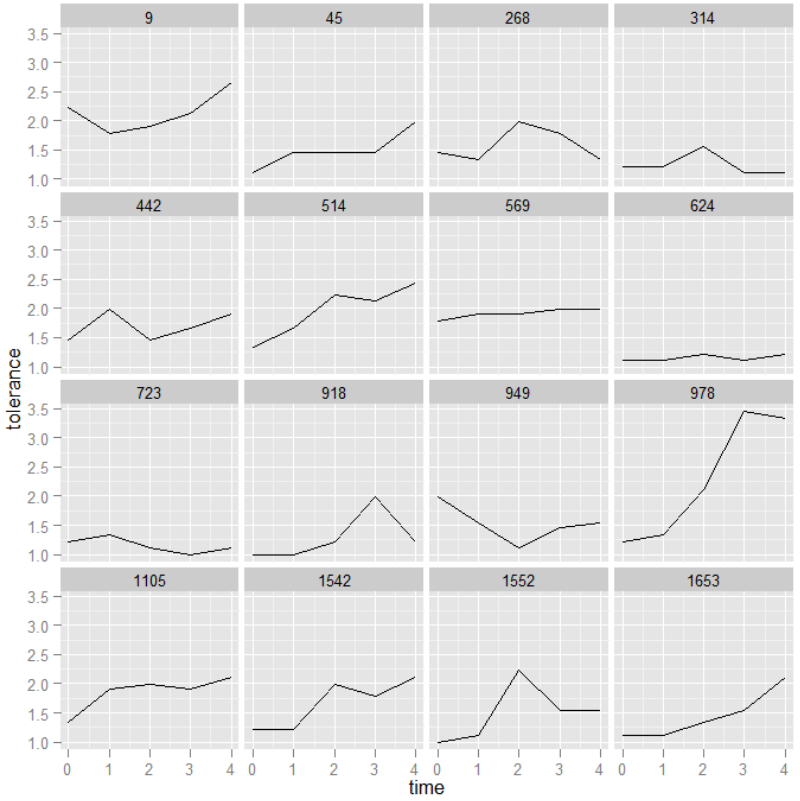


Figure 1.2: A sample of individual CD4 trajectories from the MACS data.



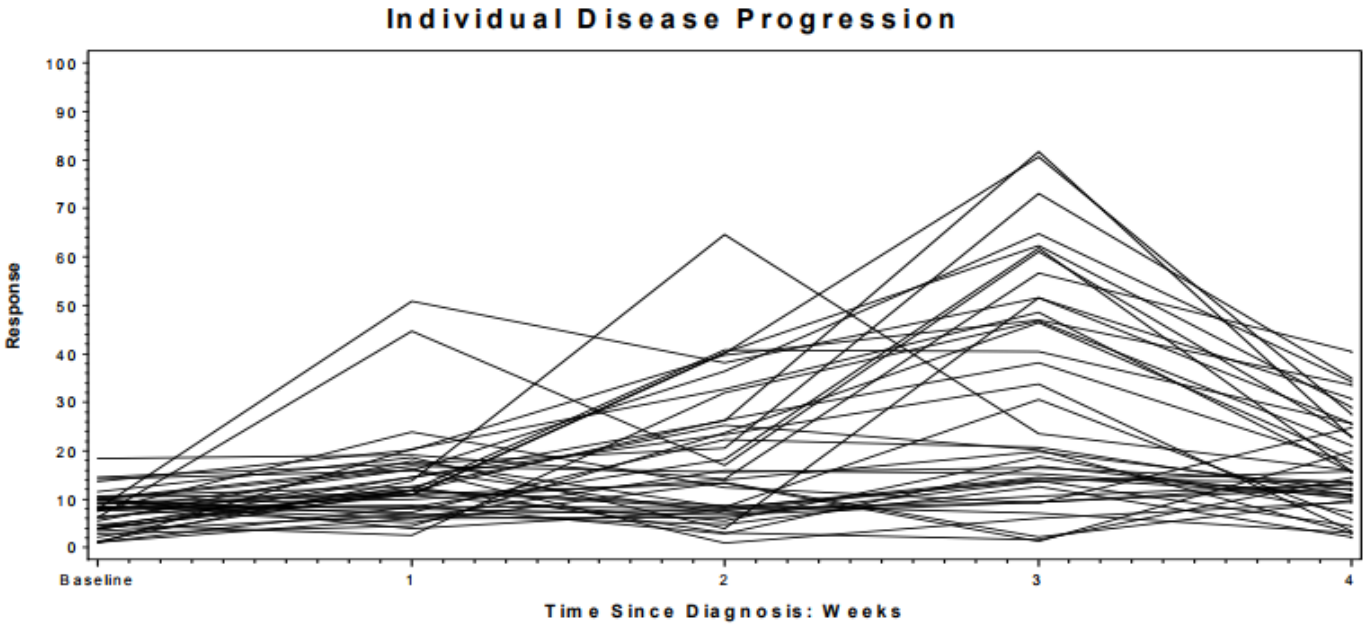
Patient trajectories

Examples
Repeated Measures/Longitudinal Plotting

```
proc gplot data=long;
      plot y*time=id / nolegend haxis=axis1 vaxis=axis2;
      symbol1 c=black i=join r=40;
run;
```

Joins the dots,

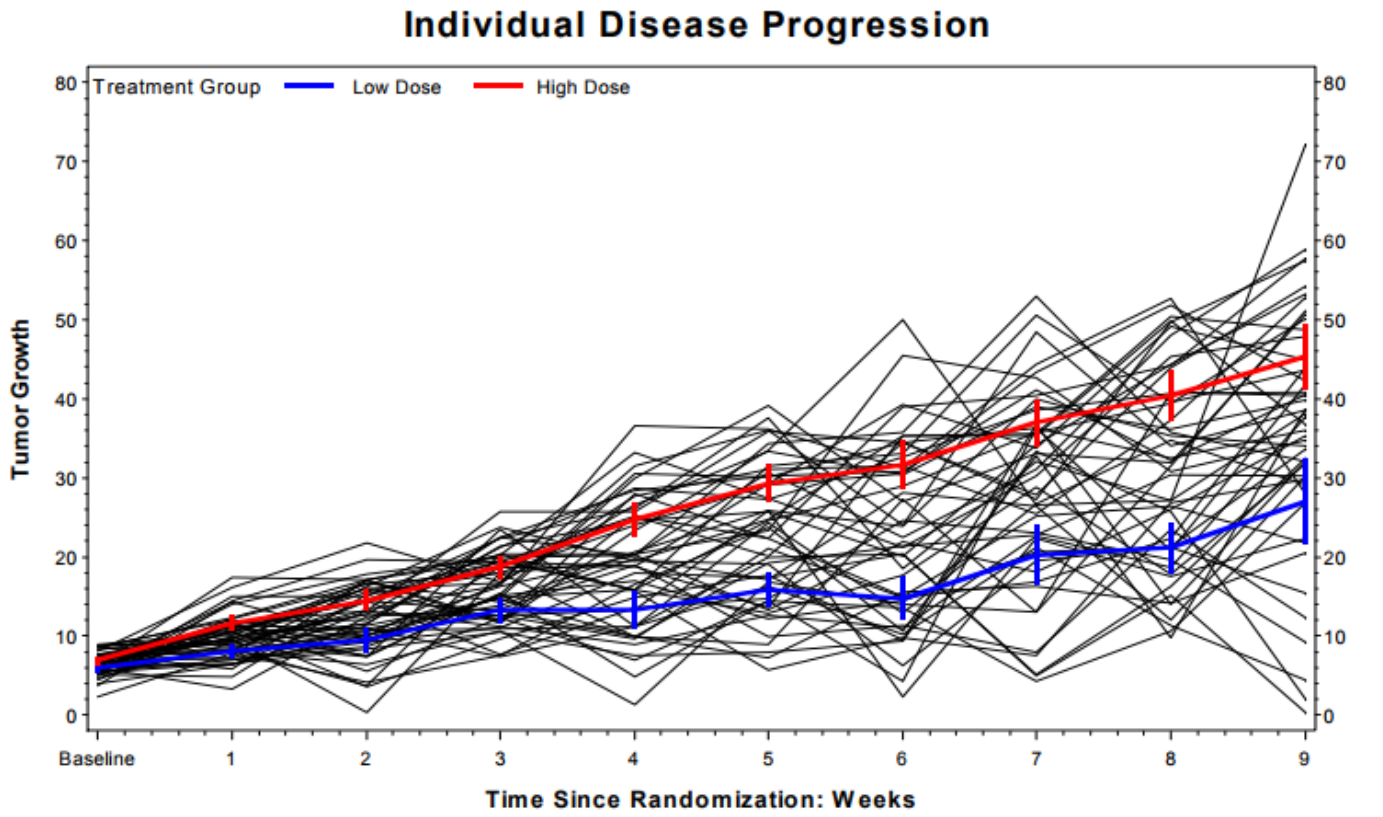
By ID



Patient trajectories

Examples

Overlay 2 plots w/ the same data



Patient trajectories

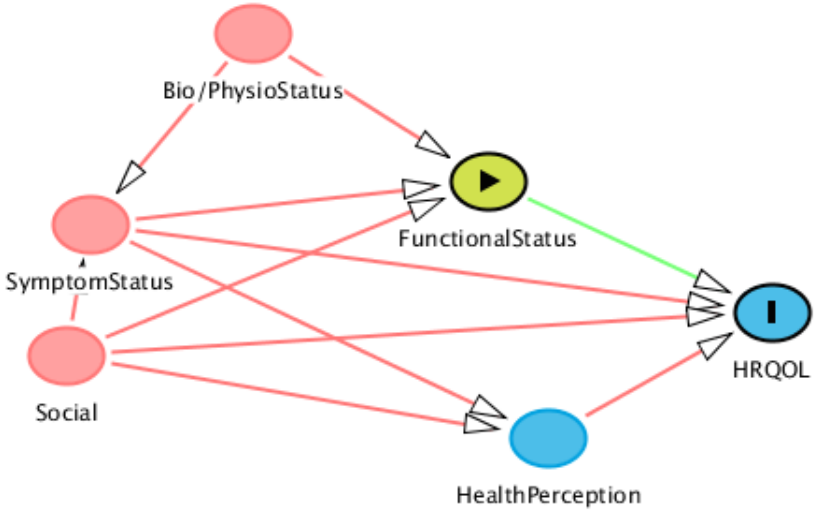
Examples

Overlay 2 plots w/ the same data

```
axis1 label=(f="arial/bo" h=1.5 "Time Since Randomization: Weeks")
      order=(1 to 10 by 1)
      value=(f="arial" h=1.2 "Baseline" "1" "2" "3" "4" "5" "6" "7" "8" "9")
      offset=(1,1);

axis2 label=(f="arial/bo" h=1.5 a=90 "Tumor Growth")
      order=(0 to 80 by 10)
      value=(f="arial" h=1.2 "0" "10" "20" "30" "40" "50" "60" "70" "80")
      offset=(1,1);

axis3 label=(f="arial/bo" h=1.5 a=90 "")
      order=(0 to 80 by 10)
      value=(f="arial" h=1.2 "0" "10" "20" "30" "40" "50" "60" "70" "80")
      offset=(1,1);

legend1 label=(f="arial" h=1.3 "Treatment Group")
        value=(f="arial" h=1.2 "Low Dose" "High Dose" )
        position=(top left inside)
        mode=protect noframe;

title "Individual Disease Progression";

proc gplot data=overlay;
      plot y*time=id /  nolegend haxis=axis1 vaxis=axis2;
      plot2 y*time=trt /  overlay legend=legend1 vaxis=axis3;

      symbol1 c=black i=join r=50 w=0.5;
      symbol2 c=blue i=stdmj l=1 w=4;
      symbol3 c=red i=stdmj l=1 w=4;
run;
```

DAGitty

- Directed acyclic graphs
- Can be used to select covariates for statistical adjustment, identify sources of bias
- can aid in this discussion among physicians and research team by providing a visual representation to discuss underlying assumptions about causal mechanism.



Resources

- ER Tufte (1983) The visual display of quantitative information. Graphics Press.
- Howard Wainer (1984) How to display data badly. The American Statistician.
- ER Tufte (1990) Envisioning information. Graphics Press.
- ER Tufte (1997) Visual explanations. Graphics Press.
- WS Cleveland (1993) Visualizing data. Hobart Press.
- WS Cleveland (1994) The elements of graphing data. CRC Press.
- A Gelman, C Pasarica, R Dodhia (2002) Let's practice what we preach: Turning tables into graphs. The American Statistician 56:121-130
- NB Robbins (2004) Creating more effective graphs. Wiley
- Nature Methods columns
- Karl Broman (2014) How to display data badly
- Karl Broman (2015) Creating effective figures and tables

Resources

- 5 Tips on designing colorblind-friendly visualizations. Tableau.
- Nathan Yau. Comparing ggplot2 and R Base Graphics.
- EJ Wagenmakers and QF Gronau. ShinyApp. A compendium of clean graphs in R.
- DAGitty.