Shutting Down the Bars: Options for Displaying and Comparing Data Distributions

John Bailer and Jim Oris
Dept. of Math. & Stat. (AJB) and Zoology (JTO)
(affiliate (AJB): Dept. of Zoology)
Ecolunch -- Jan. 30, 2007

Outline:

1. Graphical displays vs. numeric summaries

2. Default displays - tyranny of available software

3. Alternatives - hope provided by easily extended graphical tools

4. My own preferences

1. Summarizing data

“n” data values aren’t easily digested without some processing

Common numeric summaries - center (y-bar, median), spread (SD, SE)

Y-bar/SD sufficient summaries if the data are normal (describe response variability)

Y-bar/SE (describe uncertainty when estimating a population mean)

Why use graphs?

A good graphical display can capture the best of the numerical summary information plus shed light on other information (e.g. distribution shape, outliers, etc.)

Often default graphical displays don’t provide this (bar plots being the villain in this talk)

A short example - 2 groups (R you with me?)

#effectiveness of treatments for tapeworm

```
drug <- scan()
18 43 28 50 16 32 13 35 38 33 6 7

untrt <- scan()
40 54 26 63 21 37 39 23 48 58 28 39

groups <- rep(c("D","U"),
              c(length(drug), length(untrt)))
number.tapeworms <- c(drug, untrt)
```

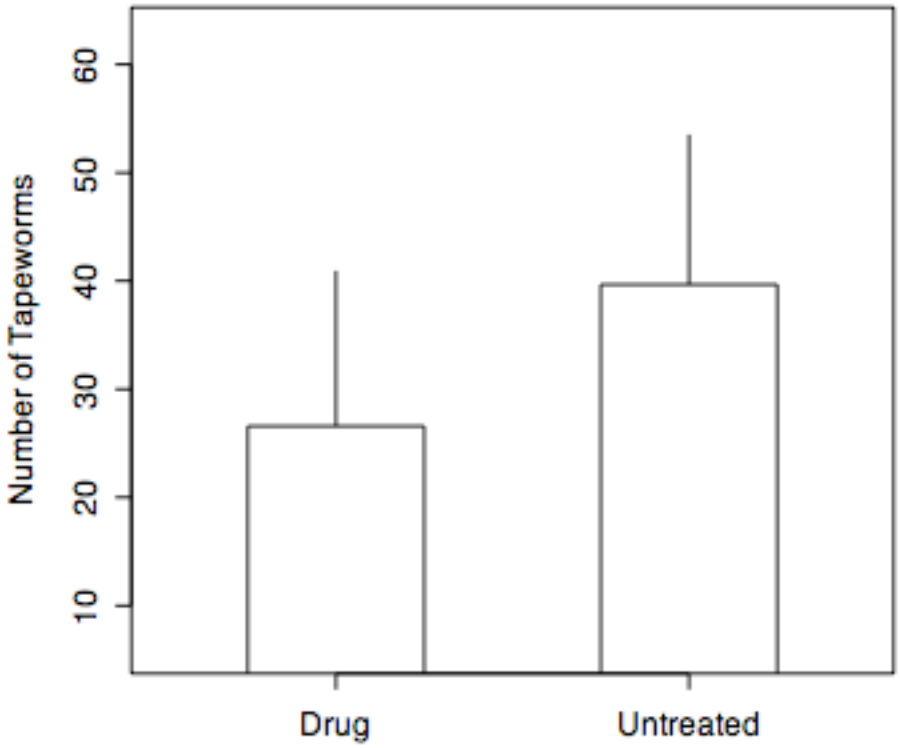
Building an analysis dataset

```
tapewm.trt.df<- data.frame(Groups=groups,
N.tapeworms=number.tapeworms)

tapewm.trt.df
   Groups N.tapeworms
1       D           18
2       D           43
…
23      U           28
24      U           39
```

2. Default summaries -

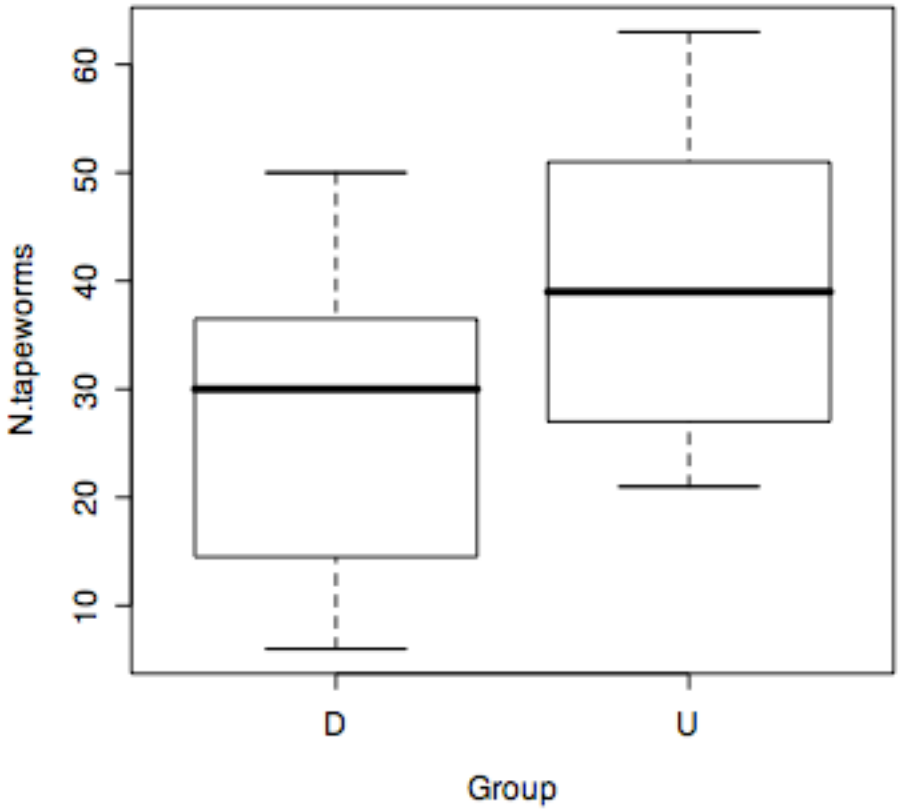
```
attach(tapewm.trt.df)

by(N.tapeworms, Groups, function(x)
c(n=length(x), ybar=mean(x),
  SD=sd(x), SE=sd(x)/sqrt(length(x))))

INDICES: D
         n      ybar         SD         SE
12.000000 26.583333 14.361934  4.145933
------------------------------------------
INDICES: U
         n      ybar         SD         SE
12.000000 39.666667 13.858593  4.000631
```

2. Default Graphical display for data of this type

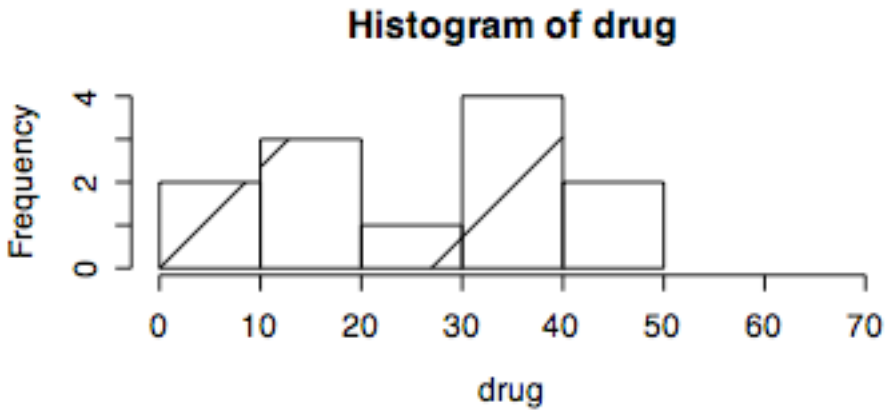


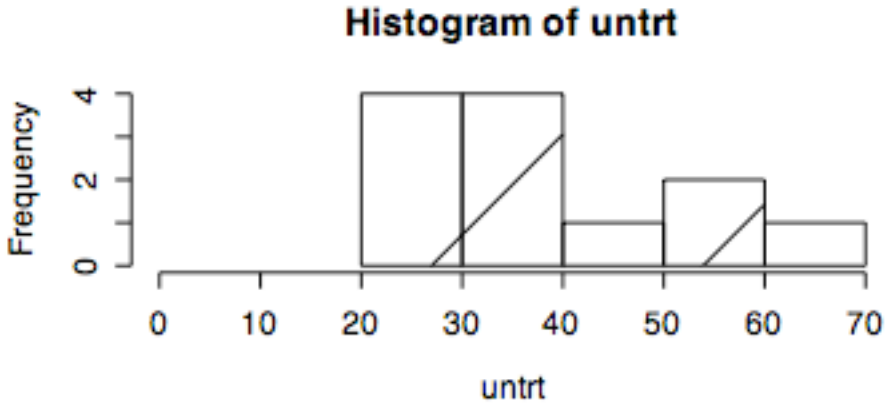
* whiskers are 1 sd here

3. Alternative Graphical displays
Boxplots

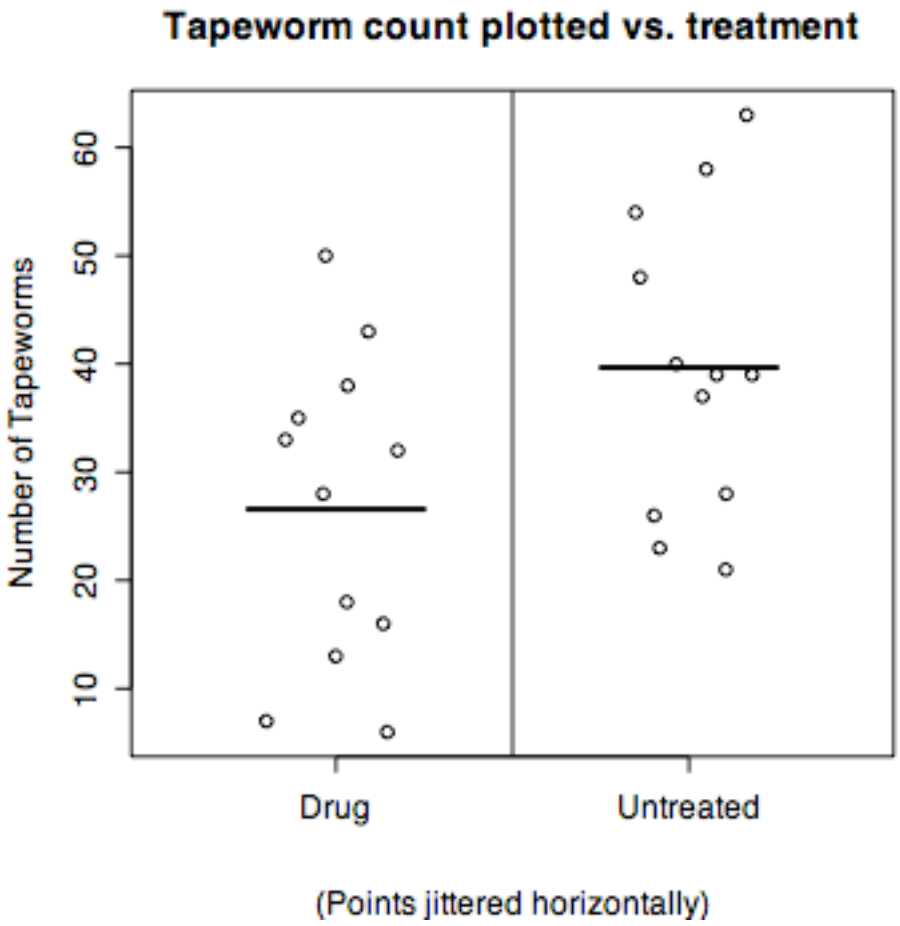


Histograms

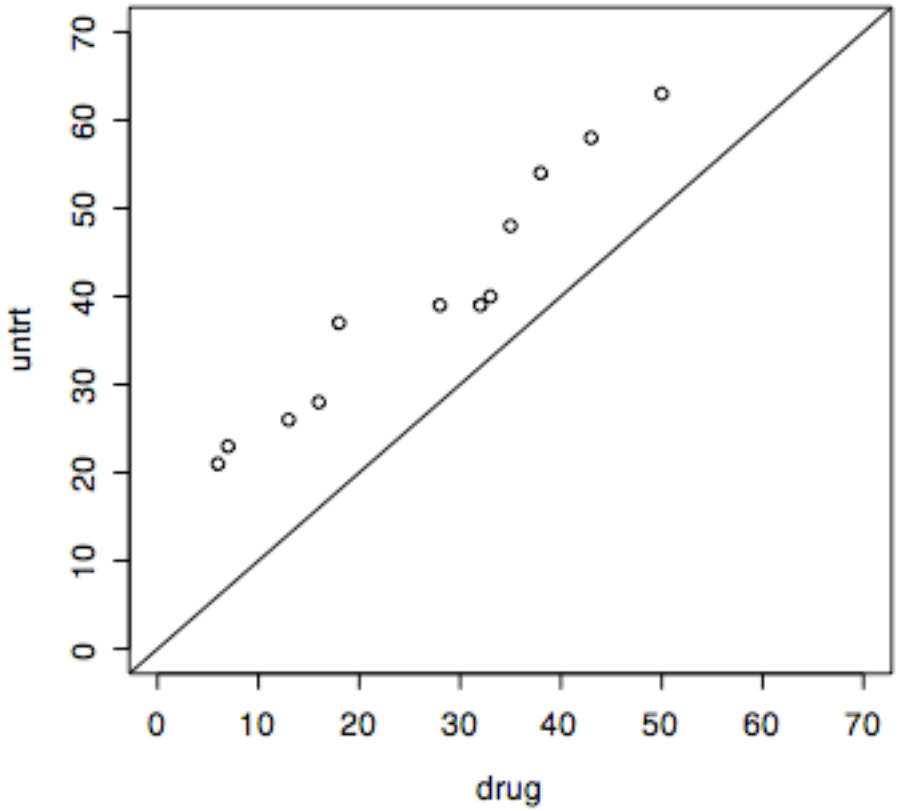

Histogram of drug
Frequency
0
2
4
0
10
20
30
40
50
60
70
drug


Histogram of untrt
Frequency
0
2
4
0
10
20
30
40
50
60
70
untrt

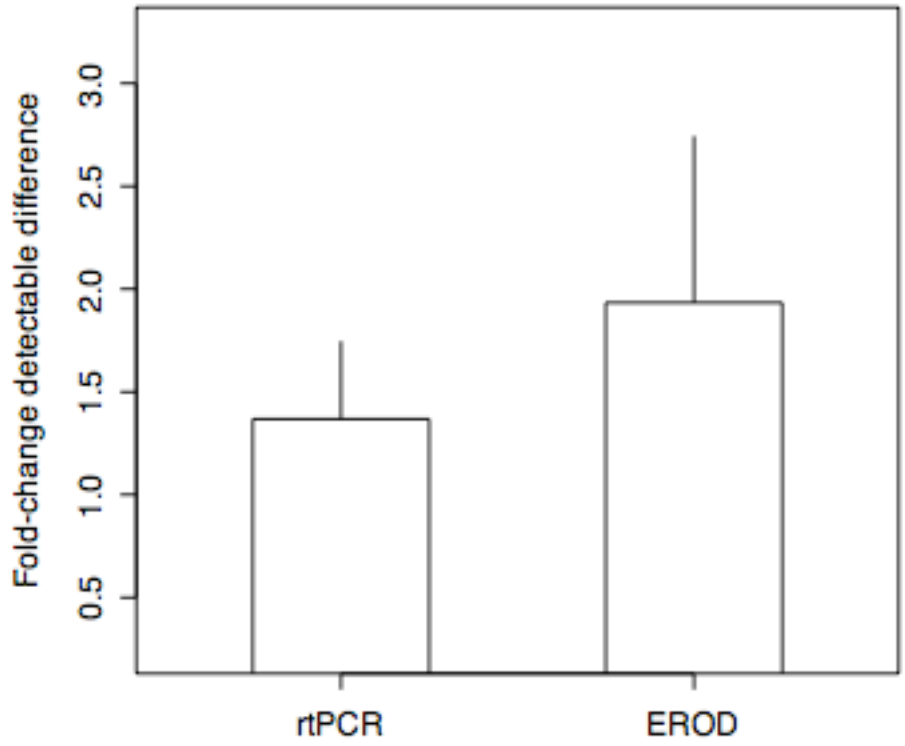
Data plot with mean value noted

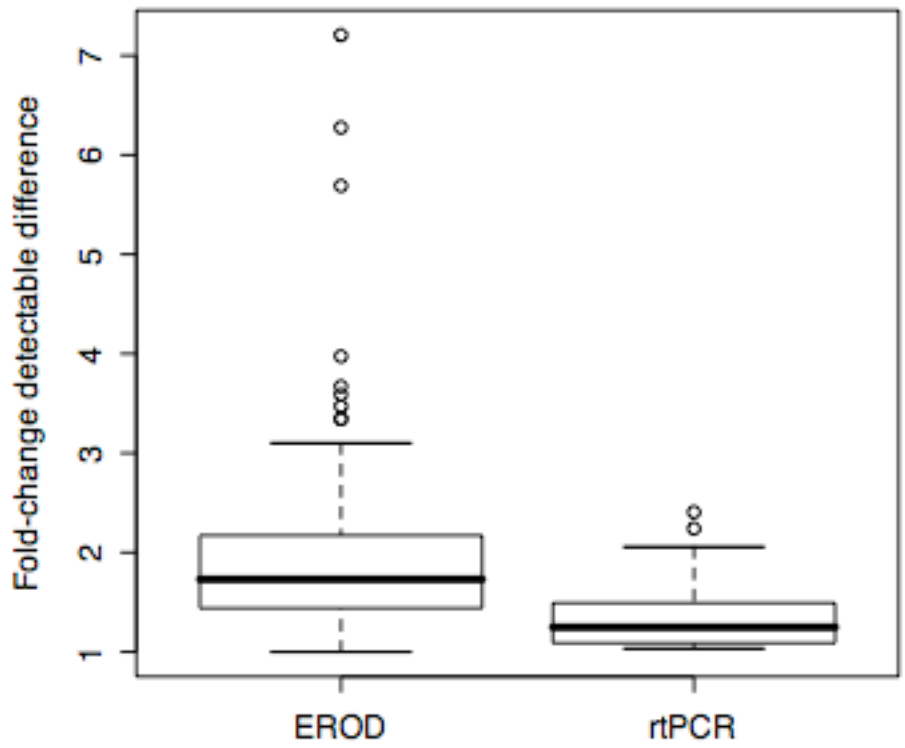


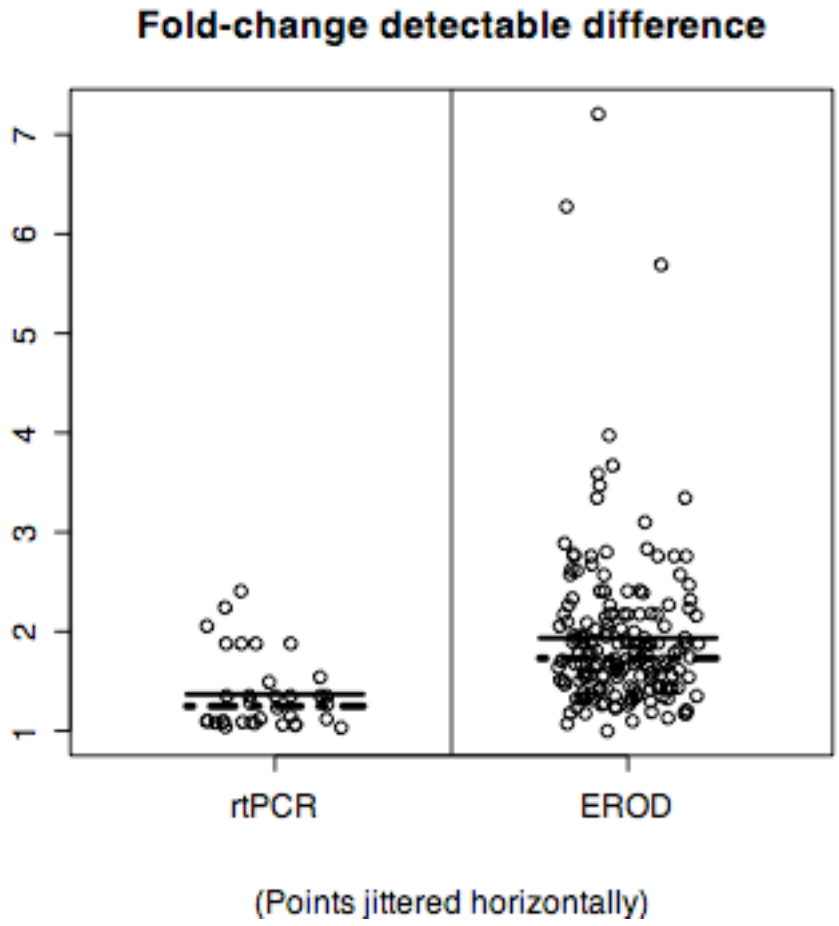
Quantiles of two distributions compared

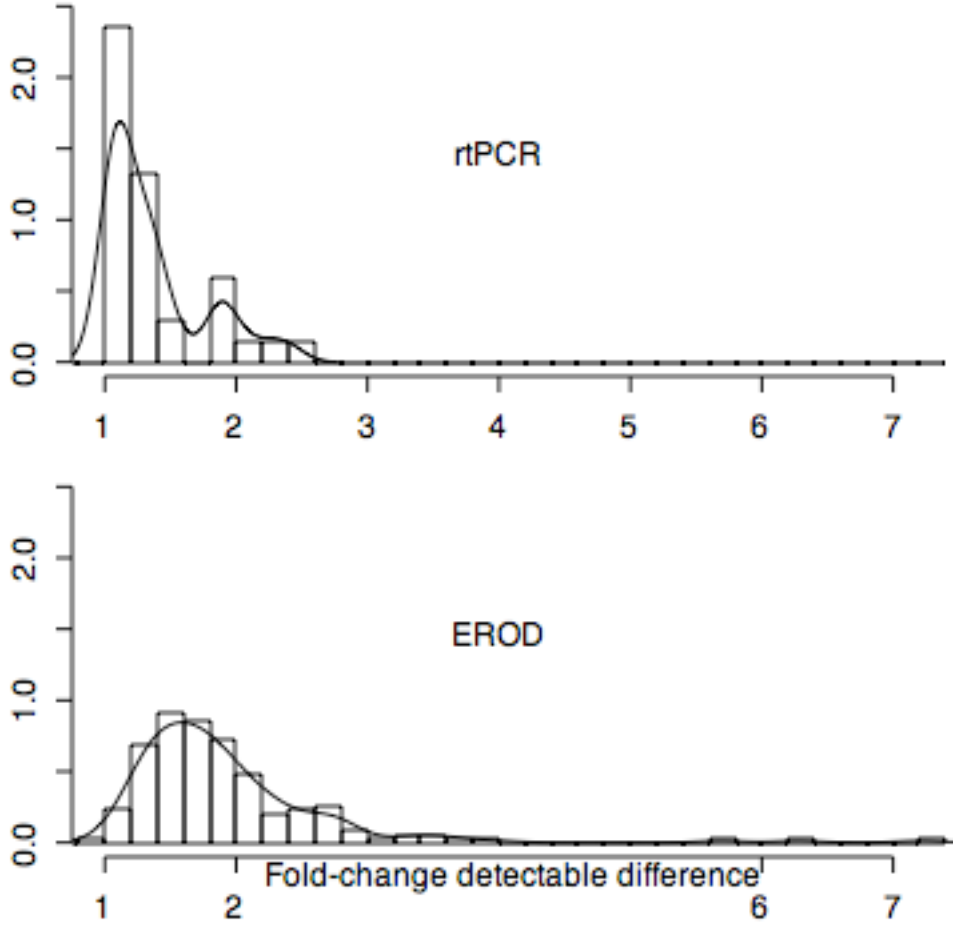


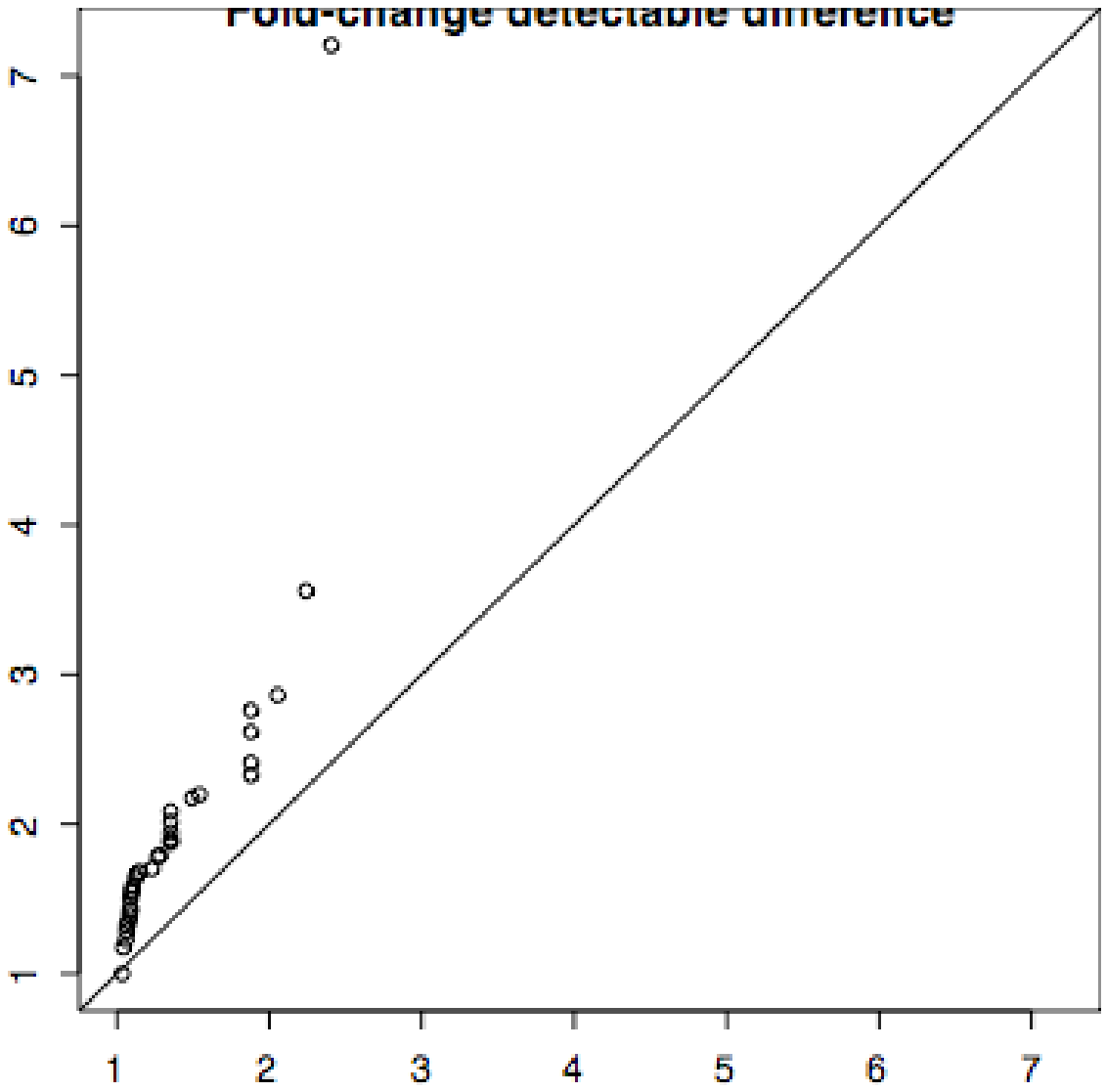
A second example (larger data sets … more skew …)




Fold-change detectable difference
1
2
3
4
5
6
7
EROD
rtPCR


Fold-change detectable difference
1
2
3
4
5
6
7
rtPCR
EROD
(Points jittered horizontally)

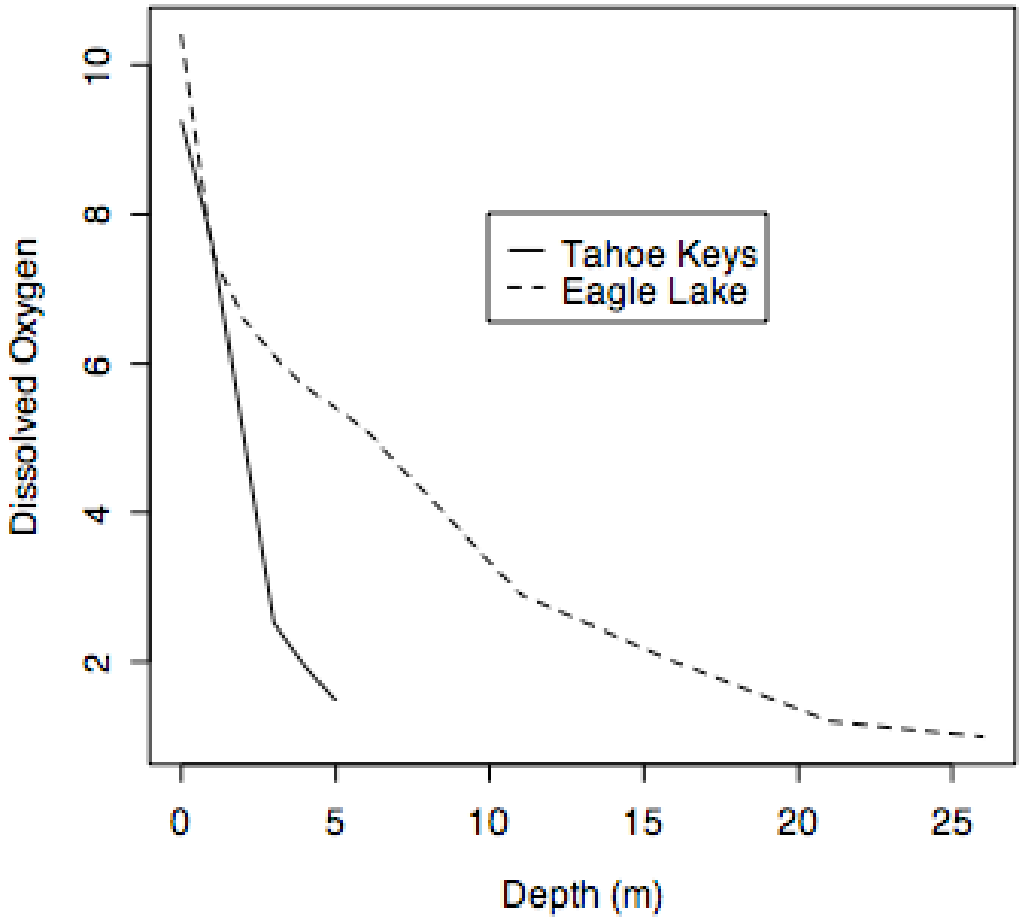

rtPCR
0.0
1.0
2.0
1
2
3
4
5
6
7
EROD
0.0
1.0
2.0
Fold-change detectable difference
1
2
6
7


Fold-change detectable difference
1
2
3
4
5
6
7

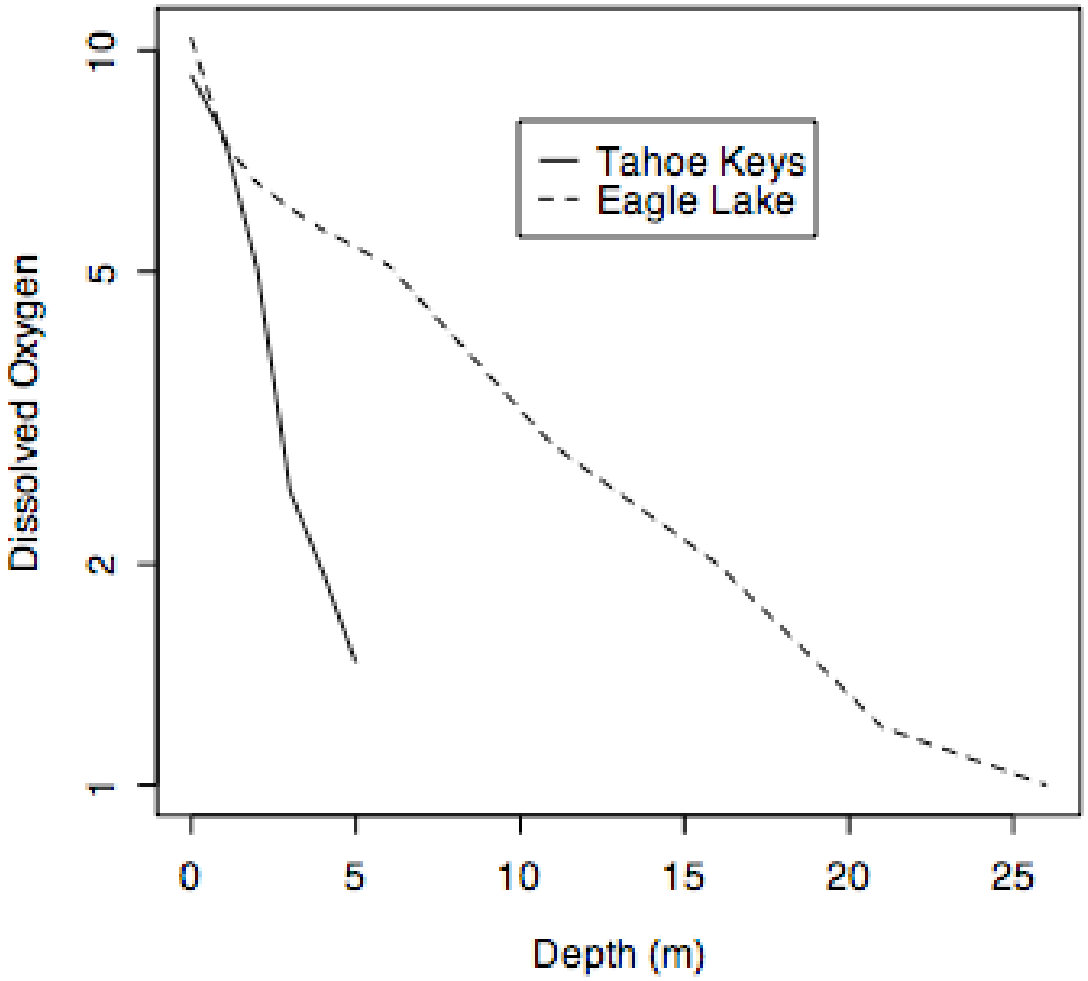
4. Summary with my preferences stated …

Graphical Display	**PROs**	**CONs**
Barplots	Familiar, journal acceptance	Doesn't display distribution shape; hides sample size info.
Boxplots	Familiar; appropriate for comparing a large # of groups; can be modified	Hides sample size info.
Histograms	Shows distribution; can superimpose density plots	Bin construction arbitrary; tough to display more than 3 groups
Data Plots	Shows all data and select summaries	Not always clear with lots of groups or obs. within groups
Quantiles	Compares the complete distribution	Limited to 2 group comparisons; Iittle more work to process

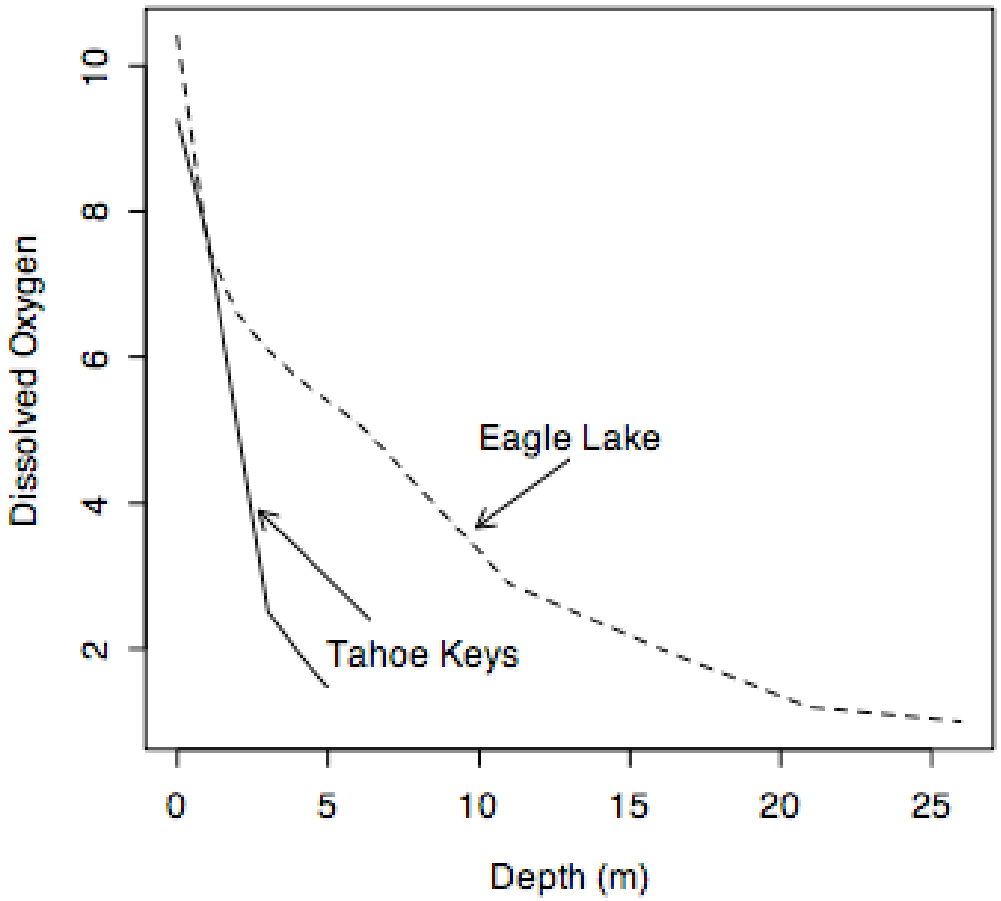
Bonus section – Death of the legend



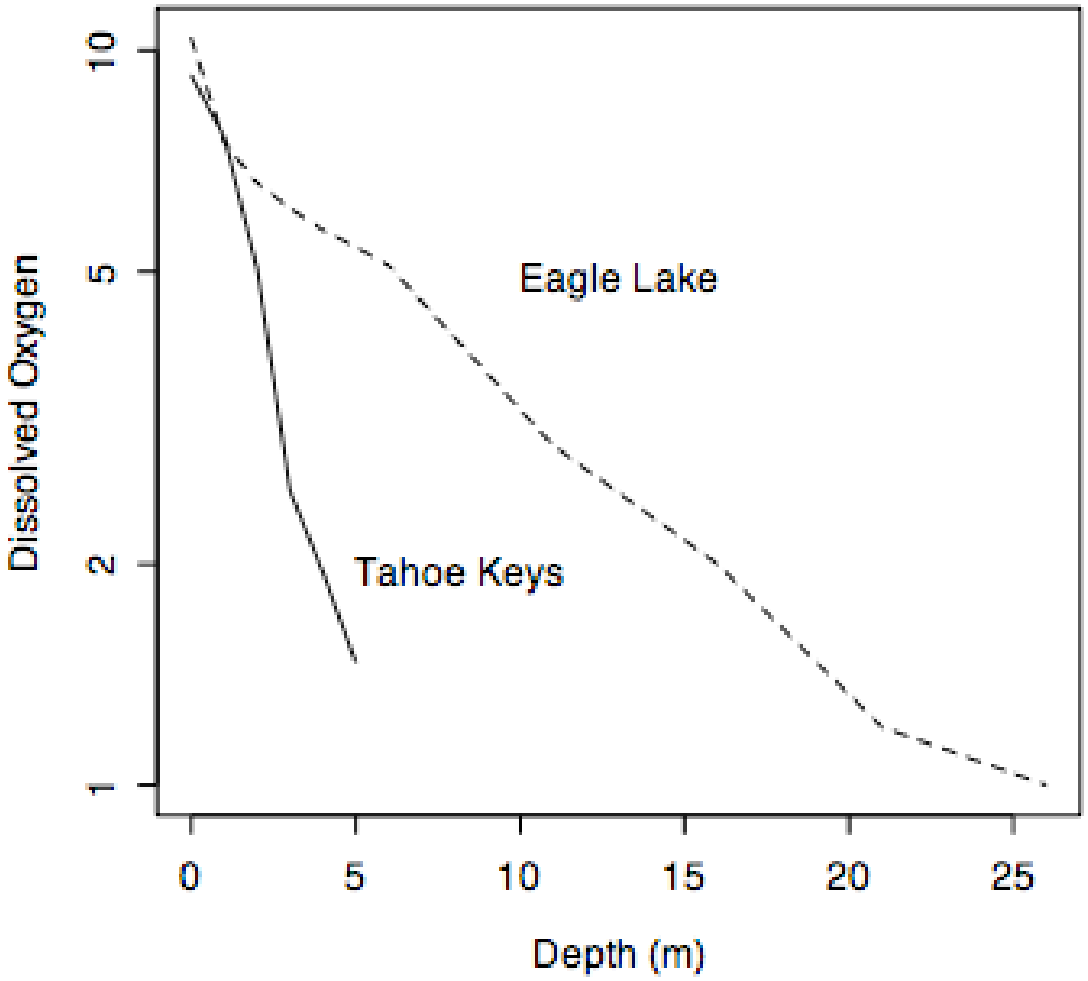
Replot with log-spacing on the y-axis



A non-legend alternative (thank you Ed Tufte)



No legend and a log DO scaling on the y-axis



Parting thoughts …

Harder work preparing figures ⇒ Easier work for the readers

Work to convey as much of the richness of your data in graphical displays (you deserve it … you collected the data!)

Appendix: R code

barplot

```
attach(ex6p3.df)
xtmp <- rep(c(1,2), c(12,12))
plot(N.tapeworms ~ xtmp, xlim=c(0.5, 2.5), xaxt="n",
type="n",
xlab="", ylab="Number of Tapeworms")
axis(1, at=c(1,2), labels=c("Drug", "Untreated"))
segments(.75, 0, .75, mean(drug))
segments(.75, mean(drug), 1.25, mean(drug))
segments(1.25, mean(drug), 1.25, 0)
segments(1.75, 0, 1.75, mean(untrt))
segments(1.75, mean(untrt), 2.25, mean(untrt))
segments(2.25, mean(untrt), 2.25, 0)
segments(1,mean(drug), 1, mean(drug) + sd(drug)) # divide
"sd" by sqrt(12) to get SE
segments(2,mean(untrt), 2, mean(untrt) + sd(untrt))
```

Boxplots

```
plot(N.tapeworms ~ factor(Groups),xlab="Group")
```

Histograms

```
par(mfrow=c(2,1))
hist(drug,density=T,xlim=c(0,70))
hist(untrt,density=T,xlim=c(0,70))
```

Plots with all data + mean

```
xtmp <- rep(c(1,2), c(12,12))
plot(N.tapeworms ~ xtmp, xlim=c(0.5, 2.5), xaxt="n",
type="n",
xlab="", ylab="Number of Tapeworms" ,
main="Tapeworm count plotted vs. treatment",
sub="(Points jittered horizontally)")
points(jitter(xtmp), N.tapeworms)
axis(1, at=c(1,2), labels=c("Drug", "Untreated"))
segments(.75, mean(drug), 1.25, mean(drug), lwd=2)
segments(1.75, mean(untrt), 2.25, mean(untrt), lwd=2)
abline(v=1.5)
```

QUANTILES of the two data sets plotted

```
qqplot(drug, untrt, xlim=c(0,70), ylim=c(0,70))
abline(a=0,b=1)
```

Code from the EROD and rtPCR

```
xtmp <- rep(c(1,2),
        c(length(detinc.rtPCR),length(detinc.EROD)))
plot(detinc ~ xtmp, xlim=c(0.5, 2.5), xaxt="n",
     type= "n", xlab="", ylab="",
     main="Fold-change detectable difference",
     sub="(Points jittered horizontally)")
points(jitter(xtmp), detinc)
axis(1, at=c(1,2), labels=c("rtPCR", "EROD"))
segments(.75, mean(detinc.rtPCR),
        1.25, mean(detinc.rtPCR), lwd=2)
segments(1.75, mean(detinc.EROD),
         2.25, mean(detinc.EROD), lwd=2)
abline(v=1.5)
```

```
# histograms with density superimposed
# along with changing plot margins
par(mfrow=c(2,1), mar=c(3,2.5,0,0))
hist(detinc.rtPCR,probability=T,
     breaks=seq(from=0,to=max(detinc)+.2,by=.2),
     xlim=range(c(detinc.rtPCR,detinc.EROD)),
     xlab="",xaxt="n",
     ylim=c(0,2.5),
     main="")
axis(1,at=1:7,paste(1:7))
text(4,1.5,"rtPCR")
lines(density(detinc.rtPCR))

hist(detinc.EROD,probability=T,
     breaks=seq(from=0,to=max(detinc)+.2,by=.2),
     xlim=range(c(detinc.rtPCR,detinc.EROD)),
     xlab="",xaxt="n",
     ylim=c(0,2.5),
     main="")
mtext("Fold-change detectable difference",side=1)
axis(1,at=c(1,2,6,7),c("1","2","6","7"))
text(4,1.5,"EROD")
lines(density(detinc.EROD))

#
# legends or not
#


lake.lst <- scan(what=list(depth=0, DO=0, lakeid=""))
```

```
    0 10.40      E    1  7.50      E    2  6.60      E
    3  6.10      E    4  5.70      E    5  5.40      E
    6  5.10      E   11  2.90      E   16  2.00      E
   21  1.20      E   26  1.00      E
    0  9.26      T    1  7.63      T    2  5.05      T
    3  2.52      T    4  1.95      T    5  1.47      T

lake.df <- as.data.frame(lake.lst)

attach(lake.df) with legend

# original DO plot
plot(depth, DO, type="n", xlab="Depth (m)", ylab="Dissolved Oxygen")
lines(depth[lakeid=="T"], DO[lakeid=="T"],lty=1)
lines(depth[lakeid=="E"], DO[lakeid=="E"],lty=2)
legend(10,8,legend=c("Tahoe Keys", "Eagle Lake"), lty=1:2)

# log(DO) plot with legend
plot(depth, DO, type="n", xlab="Depth (m)", ylab="Dissolved Oxygen", log="y")
lines(depth[lakeid=="T"], DO[lakeid=="T"],lty=1)
lines(depth[lakeid=="E"], DO[lakeid=="E"],lty=2)
legend(10,8,legend=c("Tahoe Keys", "Eagle Lake"), lty=1:2)

# original DO plot WITHOUT legend
plot(depth, DO, type="n", xlab="Depth (m)", ylab="Dissolved Oxygen")
lines(depth[lakeid=="T"], DO[lakeid=="T"],lty=1)
lines(depth[lakeid=="E"], DO[lakeid=="E"],lty=2)
text(5,2,"Tahoe Keys", adj= 0)
text(10,5,"Eagle Lake",adj= 0)
arrows(13,4.6,9.9,3.67, length=.1)
arrows(6.4, 2.4, 2.7, 3.9, length=.1)

# log DO plot WITHOUT legend
plot(depth, DO, type="n", xlab="Depth (m)", ylab="Dissolved Oxygen", log="y")
```

```
lines(depth[lakeid=="T"], DO[lakeid=="T"],lty=1)
lines(depth[lakeid=="E"], DO[lakeid=="E"],lty=2)
text(5,2,"Tahoe Keys", adj= 0)
text(10,5,"Eagle Lake",adj= 0)
# arrows(13,4.6,9.9,3.67, length=.1)
# arrows(6.4, 2.4, 2.7, 3.9, length=.1)
```