Seok-Oh Jeong, In Heok Lee, and Jay W. Rojewski
University of Georgia

The *R* Workshop

Applying the Integrated Suite of Software Facilities for Statistical Computing and Graphics

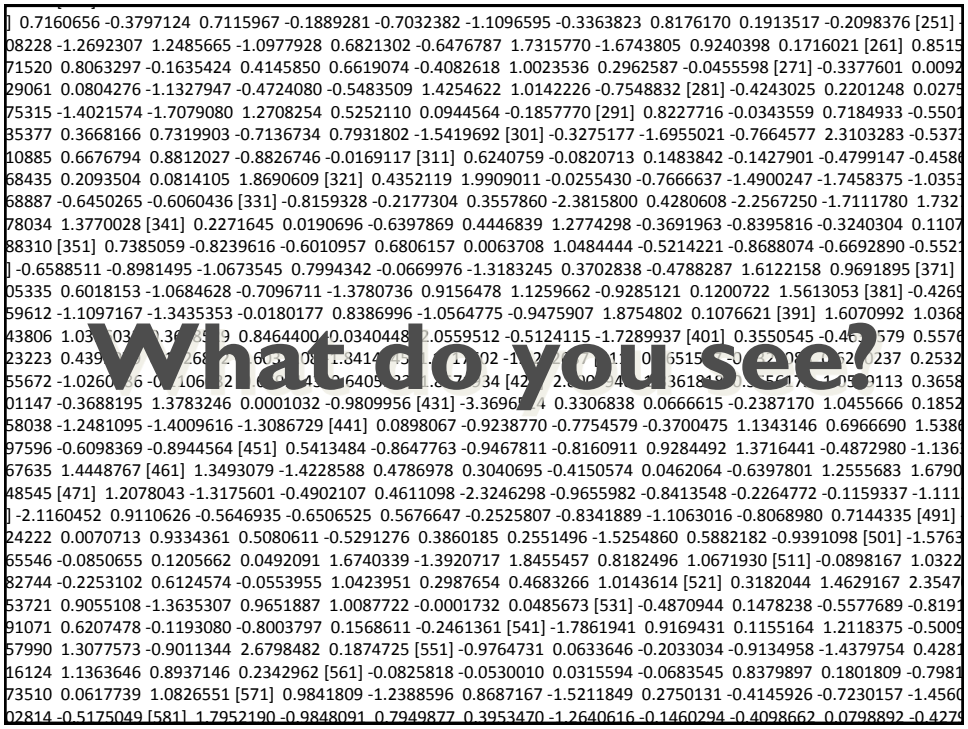
University of Georgia
Department of Workforce Education, Leadership, and Social Foundations
College of Education Research Office

January 23-January 24, 2012

January 23-January 24, 2012

6. Summarizing Data

University of Georgia
Department of Workforce Education, Leadership, and Social Foundations
College of Education Research Office



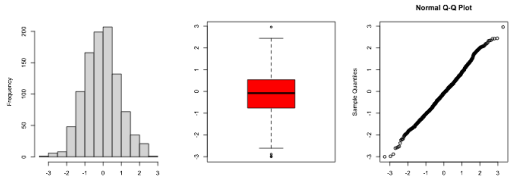
What or How to Summarize Your Data

- It depends on your research question(s)
- Do not make blind ventures with your data
- Avoid overcooking
- Use your intelligence and common sense

Summarizing Data

Distribution!

- Distributional information would be *be-all* and *end-all*
- Statistics such as mean, standard deviation, and regression coefficients stand for some partial (but important) features of the distribution
- Visualization is always the first step of your analysis



Summarizing Data

Summary Statistics

- Location (of the distribution):
  mean, median, trimmed mean...
- Scale (of the distribution):
  standard deviation, variance, range, IQR...
- Shape (of the distribution):
  skewness, kurtosis...
- Relative location (in the distribution):
  quantile, percentile, minimum, maximum...

Summarizing Data

```
library(MatchIt)
data(lalonde)
attach(lalonde)

boxplot(re78~treat)

re78.treat <- re78[treat==1]
re78.contr <- re78[treat==0]

mean(re78); sd(re78)
mean(re78.treat); sd(re78.treat)
mean(re78.contr); sd(re78.contr)

summary(lalonde)
```

Summarizing Data


```
summary.stats <- function(y)
{
x <- na.omit(y)        # Omit missing values
m <- mean(x)
s <- sd(x)
z <- (x-m)/s           # Standardization
skew <- mean(z^3)      # Skewness
kurt <- mean(z^4)      # Kurtosis
mini <- min(x)         # Minimum
maxi <- max(x)         # Maximum
q <- quantile(x, probs=c(.25, .50, .75))
res <- list(average=m, stdev=s,
            skewness=skew, kurtosis=kurt,
            q1=q[1], q2=q[2], q3=q[3],
             minimum=mini, maximum=maxi)

return(res)
}

summary.stats(re78.treat)
summary.stats(re78.contr)
```

Summarizing Data

```
table(treat, black)
table(treat, hisp)
table(treat, married)
table(treat, nodegr)
par(mfrow=c(1,2))
boxplot(age~treat, names=c("Control", "Treatment"), ylab="age")
boxplot(educ~treat, names=c("Control", "Treatment"),
 ylab="educ")
```

Summarizing Data