STAT 3355
Introduction to Data Analysis

Lecture 06: Summaries for Univariate Data I

Created by: Qiwei Li
Assistant Professor of Statistics
Presented by: Octavious Smiley
Assistant Professor of Instruction

Department of Mathematical Sciences
The University of Texas at Dallas

Last Class

- Summarize a bivariate data

	Discrete + Discrete	Discrete + Continuous	Continuous + Continuous
Numerical	`F <- xtabs(~ x + y)`		`cor(x, y)`
Graphical	`barplot(F)`		`plot(y ~ x)` `abline((lm(y ~ x)))`

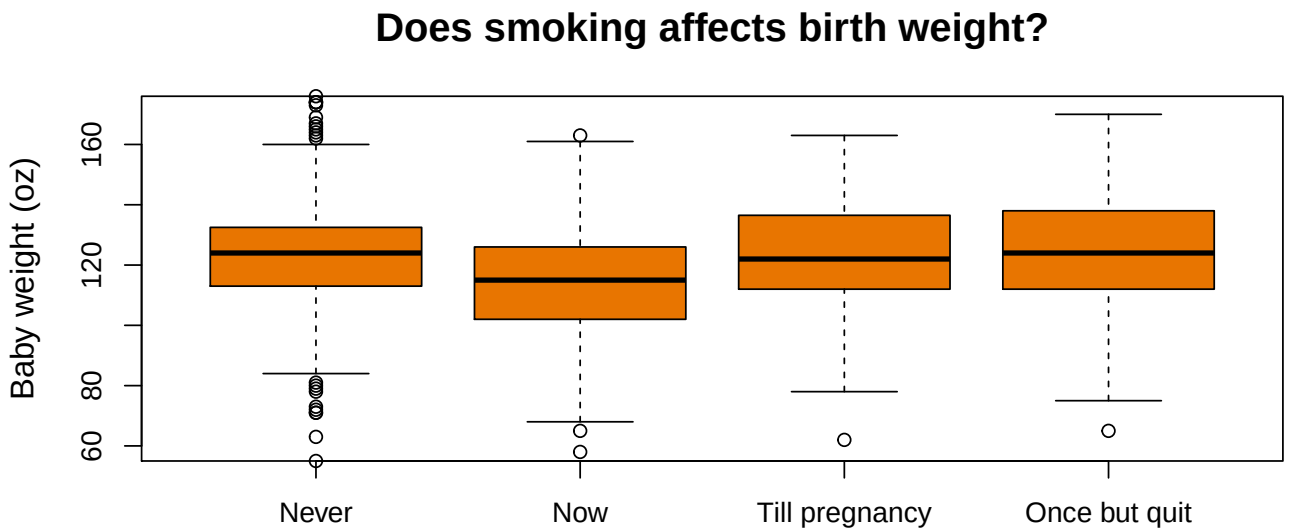
Learning Goals

- Graphical summaries for discrete and continuous data
  - Side-by-side boxplots
  - Density plots

Table of Contents

Boxplot

- Displays the five-number summary of each group side-by-side
  - One axis arranges the levels of the discrete data in some order
  - The other axis represents the value of the data
  - Each box corresponds to a unique value in the discrete data
    - IQR: Box length
    - The median: Thick line through the box
    - Outliers: points out of Whiskers



Boxplot

- Implementation in R
  - $\boldsymbol{x}$ is a numeric vector and $\boldsymbol{y}$ is an integer/factor/logical vector
    - `boxplot(`$x \sim y$`)`
  - $\boldsymbol{x}$ is a numeric variable and $\boldsymbol{y}$ is an integer/factor/logical variable in a data frame $\boldsymbol{D}$
    - `boxplot(x_name` $\sim$ `y_name, data =` D`)`

Boxplot

- Examples

```
# Load data
data("babies")

# Clean baby weight variable
babies$smoke[which(babies$smoke == 9)] <- NA
babies$smoke <- factor(babies$smoke, labels
    = c("Never", "Now", "Till pregnancy", "
    Once but quit"))

# Draw the boxplot
boxplot(wt ~ smoke, data = babies)
boxplot(wt ~ smoke, data = babies,
    horizontal = TRUE)
```

The Function `boxplot()`

- Important arguments controlling the boxes
  - `horizontal`: A logical value for the orientation of the bars
  - `range`: A number that determines how far the whiskers extend out from the box
  - `outline`: A logical value for drawing the outlines defined by the `range`
  - `col`: A vector of colors for each bar
  - `xlim` or `ylim`: A numerical vector of two values indicating the limits for the axis that represents the values

The Function `boxplot()`

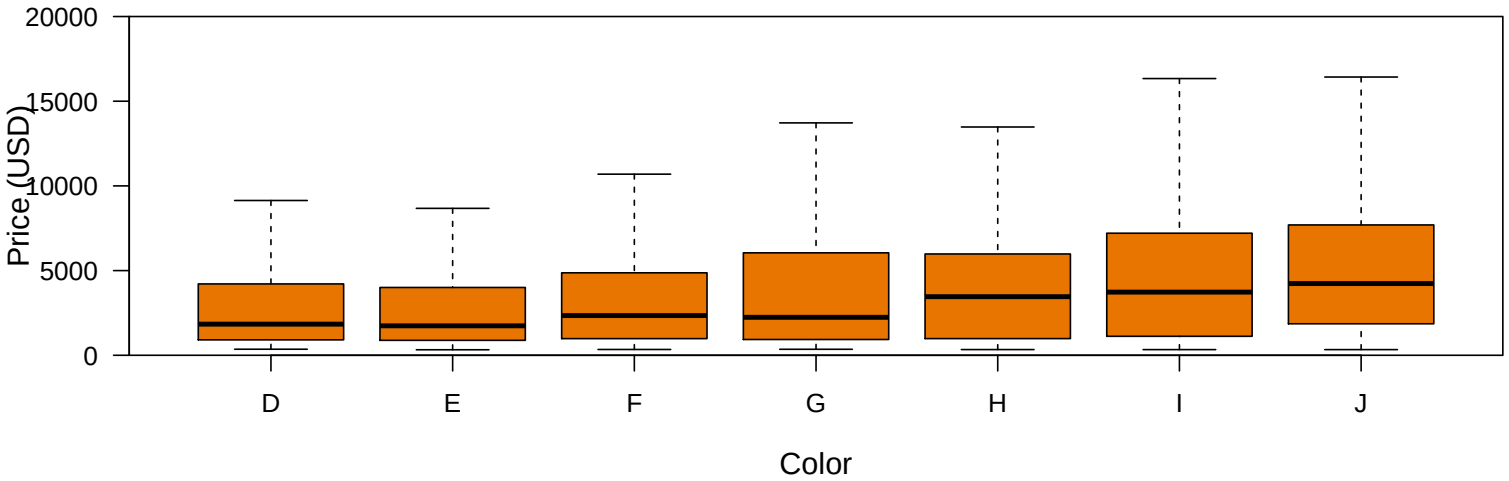
- Important arguments controlling the labels
  - `main`: A string of title
  - `xlab` and `ylab`: A string of label for the axis names
  - `las`: A numeric value of $\{0, 1, 2, 3\}$ for the orientation of axis labels

The Function `boxplot()`

- Important arguments controlling the label size
  - `cex.main`: A numeric value for the title size
  - `cex.lab`: A numeric value for the size of axis labels
  - `cex.axis`: A numeric value for the size of x axis label

Discussions

- Easy to compare the centers: The medians
- Easy to compare the spreads: The IQRs
- Easy to identify trend if the discrete data is ordinal



Your Turn

- Continue to work on the dataset `babies` in the package `UsingR`, which contains a collection of variables taken for each new mother in a Child and Health Development Study
  - Draw a side-by-side boxplot to compare birth weight between mothers whose age were greater than or equal to 35 (advanced maternal age) and those whose age were below

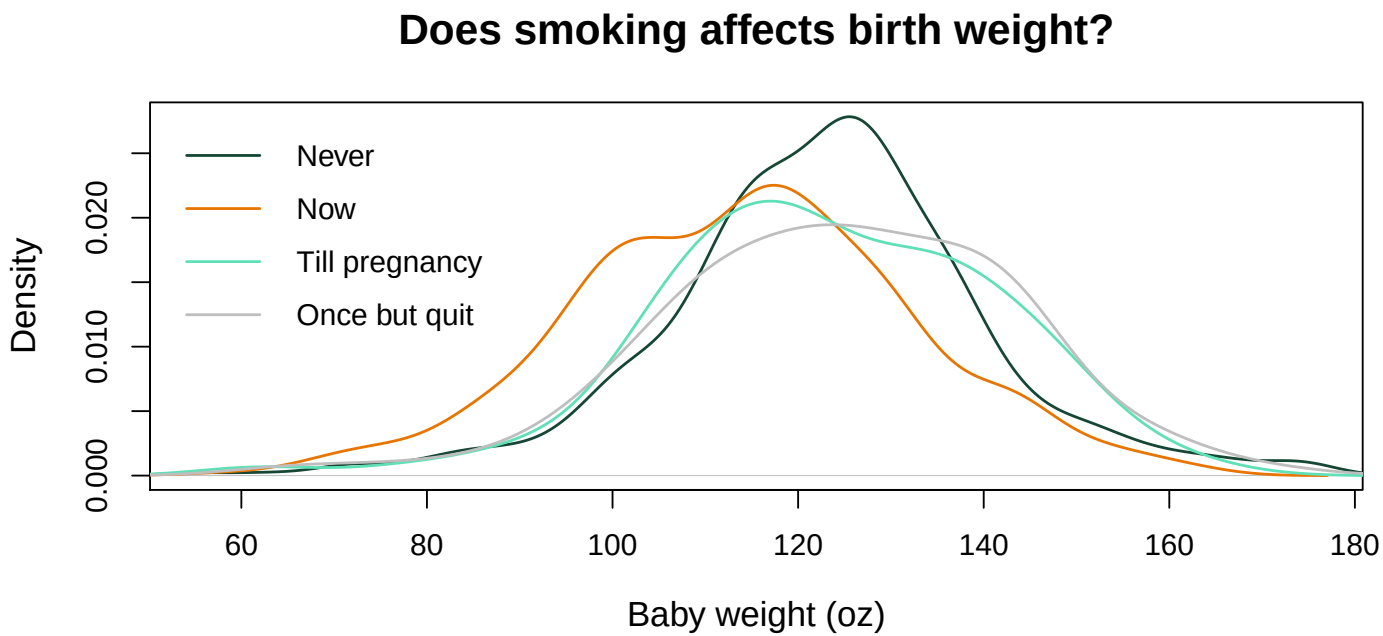
Your Turn

- Solutions

```
# Load data
data("babies")

# Clean baby age variable
babies$age[which(babies$age == 99)] <- NA

# Create a new variable indicating advanced
    maternal age
babies$ama <- babies$age >= 35

# Draw the side-by-side boxplot
boxplot(wt ~ ama, data = babies)
```

Density Plot

- Visualizes the probability distributions of a continuous data conditional on each value of the discrete data
  - x axis represents the value of the data
  - y axis represents the probability density
  - The colors indicate the levels of the discrete data in some order



Density plot

- Implementation in R
  - $\boldsymbol{x}$ is a numeric vector and $\boldsymbol{y}$ is an integer/factor/logical vector
    - `plot(density(`x`[which(`y` == )]), type = "l")` and `lines(density(`x`[which(`y` == )]))`
  - $\boldsymbol{x}$ is a numeric variable and $\boldsymbol{y}$ is an integer/factor/logical variable in a data frame $\boldsymbol{D}$
    - `plot(density(`D`$x_name[which(`D`$y_name == )]), type = "l")` and `lines(density(`D`$x_name[which(`D`$y_name == )]))`

Density Plot

- Examples

```
# Obtain the density
density_never <- density(babies$wt[which(
    babies$smoke == "Never")]);
density_now <- density(babies$wt[which(
    babies$smoke == "Now")]);
density_till <- density(babies$wt[which(
    babies$smoke == "Till pregnancy")]);
density_quit <- density(babies$wt[which(
    babies$smoke == "Once but quit")]);
```

Density Plot

- Examples

```
# Obtain the plot range
ymin <- min(density_never$y, density_now$y,
    density_till$y, density_quit$y)
ymax <- max(density_never$y, density_now$y,
    density_till$y, density_quit$y)
xmin <- min(babies$wt)
xmax <- max(babies$wt)
```

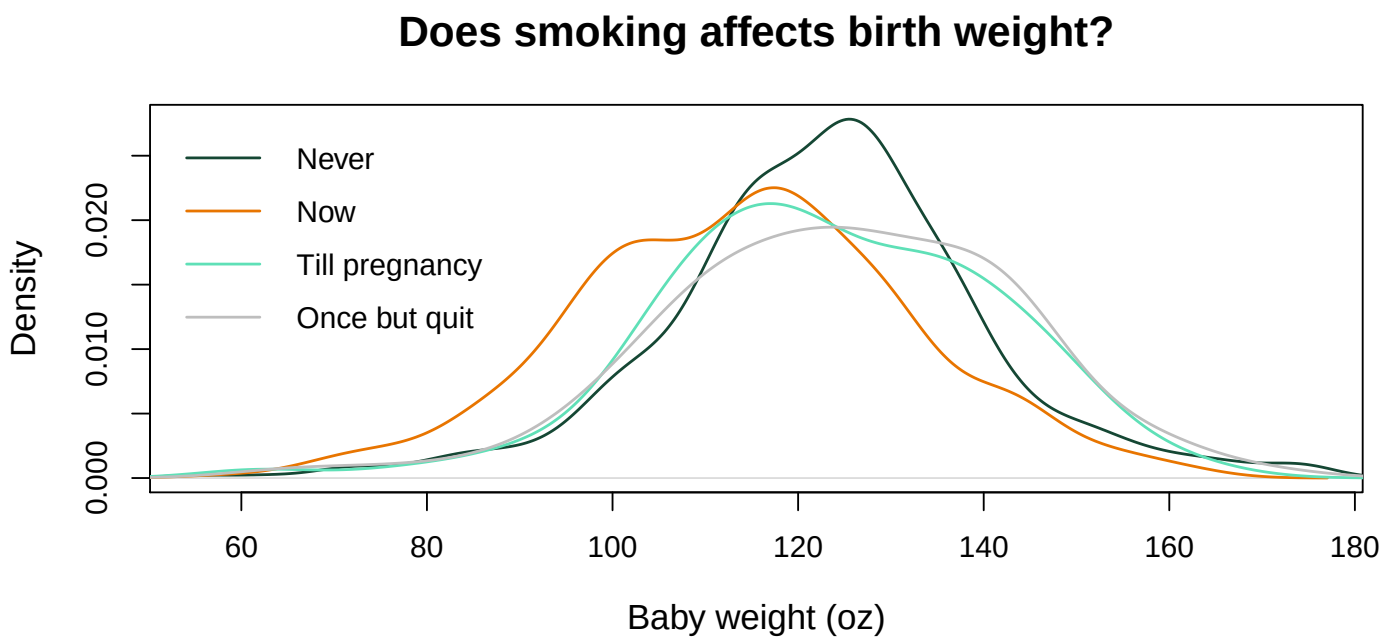
Density Plot

- Examples

```
plot(density_never, xlim = c(xmin, xmax),
    ylim = c(ymin, ymax), xlab = "Baby weight
     (oz)", main = "", lwd = 1.5, col = 1)
lines(density_now, lwd = 1.5, col = 2)
lines(density_till, lwd = 1.5, col = 3)
lines(density_quit, lwd = 1.5, col = 4)

# Add legend
legend("topleft", c("Never", "Now", "Till
    pregnancy", "Once but quit"), col = 1:4,
    lwd = rep(1.5, 4), lty = c(1, 1, 1, 1),
    bty = 'n')
```

Discussions

- Easy to compare the centers: The means and the modes
- Easy to compare the spreads
- Easy to identify the shapes: Symmetry, skewness, and multi-modality



After-class Reading

- *Using R for Introductory Statistics (1st Ed.)* by John Verzani
- Chapter 3 Bivariate data
  - Section 3.2 Comparing independent samples
    - Subsection 3.2.1 Side-by-side boxplots
    - Subsection 3.2.2 Densityplots

After-class Reading

- *Using R for Introductory Statistics (2nd Ed.)* by John Verzani
- Chapter 3 Bivariate data
  - Section 3.1 Independent samples