

# **R Programming and Data Analysis**

## **Graphics**

# Introduction

1. Traditional graphics
2. Plotting with ggplot2
3. Reproducible research & documentation in R
4. Shiny

# Traditional Graphics

# Traditional Graphics

**... Demonstration ...**  
(See `traditional.Rmd`)

# ggplot2

# ggplot2

- ggplot2 is a popular graphics package whose syntax is based on a “grammar for graphics”.
- Components of a graphic are
  1. data frames
  2. aesthetics
  3. geoms
  4. facets
  5. statistical transformations
  6. scales
  7. coordinate systems
- For further reference
  1. Video tutorial by Roger Peng: <https://youtu.be/HeqHMM4ziXA>.
  2. Book (Wickham, 2010).
  3. ggplot2 Cheat Sheet

# Traditional Graphics

**... Demonstration ...**  
(See `ggplot2.Rmd`)

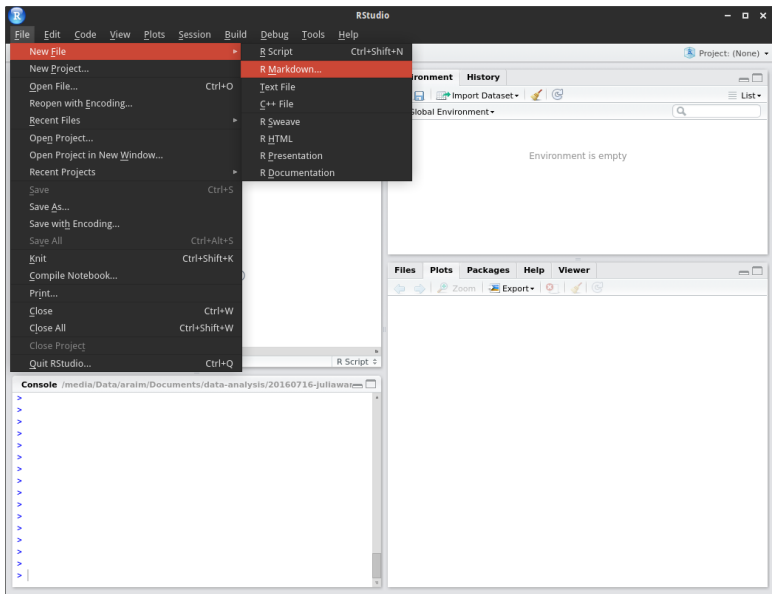
# Reproducible Research & Documentation



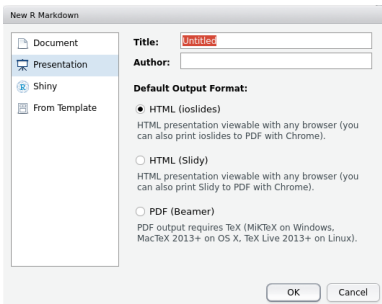
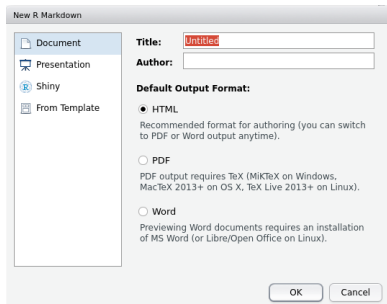
# Reproducible Research & Documentation

- **Reproducible research** is an idea which promotes releasing code and data from an analysis, not only the results.
- This allows others to verify details of your analysis and try alternative methods.
- There is a documentation system integrated into Rstudio.
- Embed chunks of R code into your document and update dynamically when document is compiled, via `knitr` package (Xie, 2015).
- R Markdown (<http://rmarkdown.rstudio.com>) is a Wiki-like language used for authoring many kinds of documents.
  1. Reports: HTML, PDF, Word
  2. Slides: HTML, Beamer
  3. Books and technical documents (Xie, 2016)
  4. Blog posts (Xie and Hill, 2017)
  5. And more: <http://rmarkdown.rstudio.com/formats.html>
- What to do about long computations embedded in document?

# Reproducible Research & Documentation



# Reproducible Research & Documentation



```

---
title: "Example Markdown Document"
author: "Andrew R"
date: "July 24, 2016"
output: pdf_document
---

# Introduction #
Consider a random sample from the normal distribution

$$X_1, \dots, X_{200} \sim \text{N}(1, 2^2).$$


Let's take a sample.
```{r}
x <- rnorm(200, 1, 2)
summary(x)
```

Here is a histogram.
```{r, echo=FALSE, out.width="2in", out.height="2in",
  fig.align="center"}
hist(x)
```

The mean of my sample was r mean(x) and the standard deviation was
r sd(x). Recall that  $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu$ 
and  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \rightarrow \sigma^2$ 
almost surely, as  $n \rightarrow \infty$ .

Here is some code for display only.
```{r, eval=FALSE}
cat("Hello World")
```

```

# Reproducible Research & Documentation

The screenshot displays the RStudio environment with three main panes:

- Source Editor:** Contains an R script named `doc.Rmd`. The script includes comments and code for generating a histogram and calculating sample statistics. It uses `knitr` syntax for embedding LaTeX and HTML.
- Environment:** Shows the current environment is empty.
- Console:** Displays the output of the R script, including the histogram and the calculated mean and standard deviation.
- Viewer:** Shows the rendered output of the R Markdown document, which includes the histogram and the calculated statistics.

```
## Introduction ##
## Consider a random sample from the normal distribution
##
## Let's take a sample.
##
## Here is a histogram.
##
## The mean of my sample was 'r mean(x)' and the standard deviation was 'r
## sd(x)'. Recall that  $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu$  and
##  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \rightarrow \sigma^2$  almost
## surely, as  $n \rightarrow \infty$ .
##
## Here is some code for display only.
##
```

Console Output:

```
label: unnamed-chunk-3 (with options)
list of 1
 $ eval: logi FALSE

/usr/lib/rstudio/bin/pandoc/pandoc +RTS -K512M -RTS doc.utf8.md --to latex --fr
om markdown+autolink_bare_uris+asciit_identifiers+tex_math_single_backslash -o
tput doc.pdf --template /home/arain/R/x86_64-pc-linux-gnu-library/3.3/rmarkdown
```

Viewer Output:

cropping doc\_files/figure-latex/unnamed-chunk-2-1.pdf  
PDFCROP 1.38, 2012/11/02 - Copyright (c) 2002-2012 by Heiko Oberdiek.  
=> 1 page written on 'doc\_files/figure-latex/unnamed-chunk-2-1.pdf'.  
|.....| 83%  
inline R code fragments  
|.....| 100%  
label: unnamed-chunk-3 (with options)  
list of 1  
\$ eval: logi FALSE

Markdown Quick Reference:

R Markdown is an easy-to-write plain text format for creating dynamic documents and reports. See [Using R Markdown](#) to learn more.

**Emphasis**

- `*italic*` `**bold**`
- `_italic_` `__bold__`

**Headers**

- `# Header 1`
- `## Header 2`
- `### Header 3`

**Lists**

Unordered List

- `* Item 1`
- `* Item 2`
- `+ Item 2a`
- `+ Item 2b`

Ordered List

# Reproducible Research & Documentation

doc.pdf — Example Markdown Document

File Edit View Go Bookmarks Help

1 of 1

80.21%

## Example Markdown Document

*Andrew R*  
*July 24, 2016*

### Introduction

Consider a random sample from the normal distribution

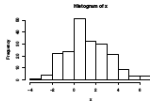
$$X_1, \dots, X_{200} \sim N(1, 2^2).$$

Let's take a sample.

```
x <- rnorm(200, 1, 2)
summary(x)
```

| ## | Min.     | 1st Qu.  | Median  | Mean    | 3rd Qu. | Max.    |
|----|----------|----------|---------|---------|---------|---------|
| ## | -3.37500 | -0.02034 | 0.97130 | 1.20000 | 2.49800 | 6.22000 |

Here is a histogram.



The mean of my sample was 1.1999757 and the standard deviation was 1.8745499. Recall that  $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu$  and  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \rightarrow \sigma^2$  almost surely, as  $n \rightarrow \infty$ .

Here is some code for display only.

```
cat("Hello World")
```

```

---
title: "Example Slides"
author: "Andrew R"
output: beamer_presentation
fontsize: 10pt
bibliography: references.bib
---

## R Markdown
This is a sample presentation using R Markdown, with R code knitted in.

- Draw a random sample from normal.
- Compute a summary.
- Plot a histogram.

See the book about knitr by @Xie2015.

## Slide with R Code and Output {.smaller}
```{r}
x <- rnorm(200, 1, 2)
summary(x)
```

```{r, out.height="2in", out.width="2in", fig.align='center'}
hist(x)
```

## References

```





## Example Slides

Andrew R

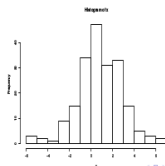
Navigation icons

## Slide with R Code and Output

```
x <- rnorm(200, 1, 2)
summary(x)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5.6240 -0.2962   0.7280   0.8635  2.3890   6.3140
```

```
hist(x)
```



Navigation icons

## R Markdown

This is a sample presentation using R Markdown, with R code knitted in.

- ▶ Draw a random sample from normal.
- ▶ Compute a summary.
- ▶ Plot a histogram.

See the book about knitr by Xie (2015).

Navigation icons

## References

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Chapman; Hall/CRC.

Navigation icons

# Shiny

# Shiny

- Shiny (<http://shiny.rstudio.com>) is a web application framework for R.
- Develop and run applications in RStudio for yourself.
- Deploy applications to the web for public use.
  1. Via <http://www.shinyapps.io> hosting service.
  2. Or by deploying a Shiny Server.
- In-depth tutorials are available at <http://shiny.rstudio.com/tutorial>.

# Beta-Binomial Distribution

- The Beta-Binomial (BB) distribution is an extension of Binomial which allows for extra variation.
- $Z \sim \text{Binomial}(m, p)$  measures the number of successes out of  $m$  independent success/failure trials, each having success probability  $p$ .
- Suppose

$$Y \sim \text{Binomial}(m, p)$$

$$p \sim \text{Beta}(\pi\rho^2(1 - \rho^2), (1 - \pi)\rho^2(1 - \rho^2)),$$

where  $a = \pi(1 - \rho^2)/\rho^2$  and  $b = (1 - \pi)(1 - \rho^2)/\rho^2$ .

- Marginally,  $Y \stackrel{\text{iid}}{\sim} \text{BB}(\pi, \rho)$ , with density

$$f(y \mid m, \pi, \rho) = \frac{\Gamma(m+1)}{\Gamma(y+1)\Gamma(m-y+1)} \frac{\Gamma(a+y)\Gamma(b+m-y)\Gamma(a+b)}{\Gamma(a+b+m)\Gamma(a)\Gamma(b)},$$

$$E(Y) = m\pi$$

$$\text{Var}(Y) = m\pi(1 - \pi)\{1 + \rho(m - 1)\}$$

# Beta-Binomial Shiny App

**... Demonstration ...**  
(See betabin-shiny subdirectory)

# Dirichlet Distribution

- The Dirichlet distribution models probabilities  $\mathbf{X} = (X_1, \dots, X_k)$  that sum to 1 (“compositional data”).
- It is an extension of the Beta distribution (where  $k = 2$ ).
- We can write  $\mathbf{X} \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$ , if data is drawn from Dirichlet with density

$$f(\mathbf{x}) = \frac{x_1^{\alpha_1-1} \cdots x_k^{\alpha_k-1}}{B(\alpha_1, \dots, \alpha_k)}, \mathbf{x} \in \mathcal{S}^k,$$

where

$$B(\alpha_1, \dots, \alpha_k) = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}{\Gamma(\alpha_1 + \cdots + \alpha_k)}$$
$$\mathcal{S}^k = \left\{ (x_1, \dots, x_k) : x_j \in (0, 1), \sum_{j=1}^k x_j = 1 \right\}.$$

# Shiny

## Dirichlet Shiny App

**... Demonstration ...**  
(See `dirichlet-shiny` subdirectory)

# References I

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2010.

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, 2nd edition, 2015.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, 2016.

Yihui Xie and Alison Presmanes Hill. *blogdown: Creating Websites with R Markdown*. Chapman and Hall/CRC, 2017.