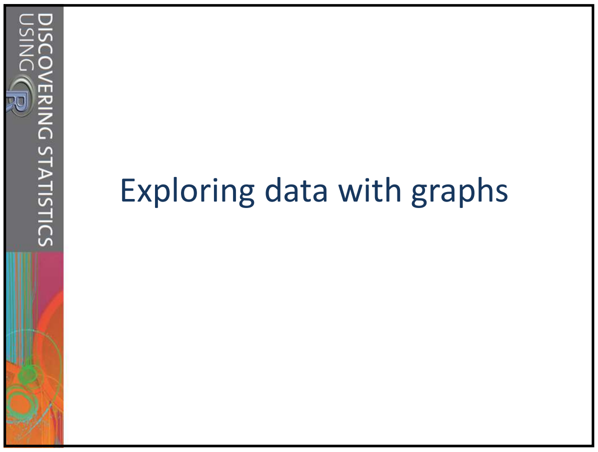
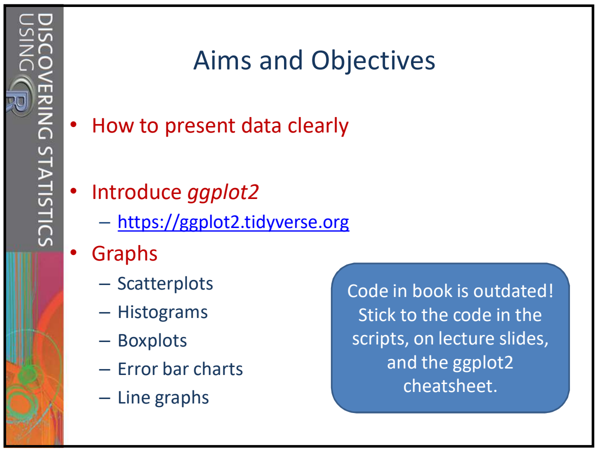

DISCOVERING STATISTICS USING R

Exploring data with graphs

1

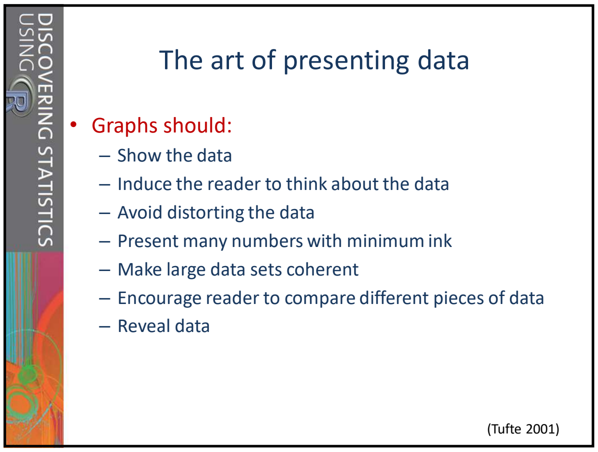

DISCOVERING STATISTICS USING R

Aims and Objectives

- How to present data clearly
- Introduce *ggplot2*
  - https://ggplot2.tidyverse.org
- Graphs
  - Scatterplots
  - Histograms
  - Boxplots
  - Error bar charts
  - Line graphs

Code in book is outdated! Stick to the code in the scripts, on lecture slides, and the ggplot2 cheatsheet.

2

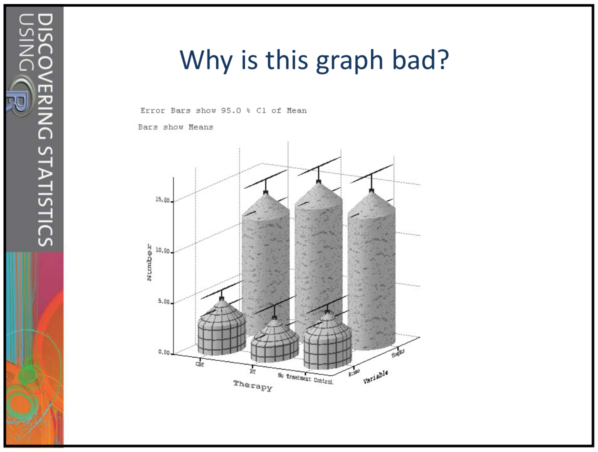

DISCOVERING STATISTICS USING R

The art of presenting data

- Graphs should:
  - Show the data
  - Induce the reader to think about the data
  - Avoid distorting the data
  - Present many numbers with minimum ink
  - Make large data sets coherent
  - Encourage reader to compare different pieces of data
  - Reveal data

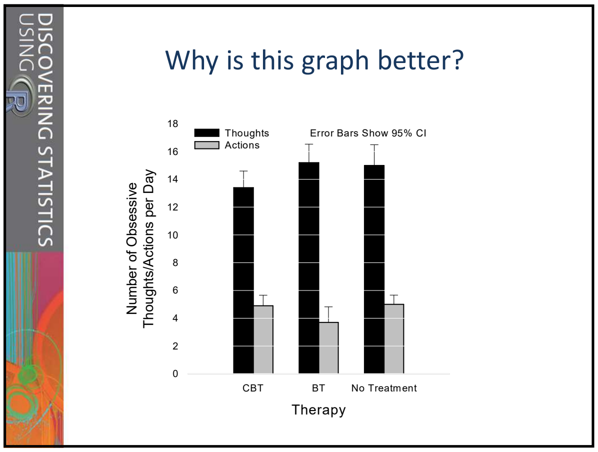
(Tufte 2001)

3



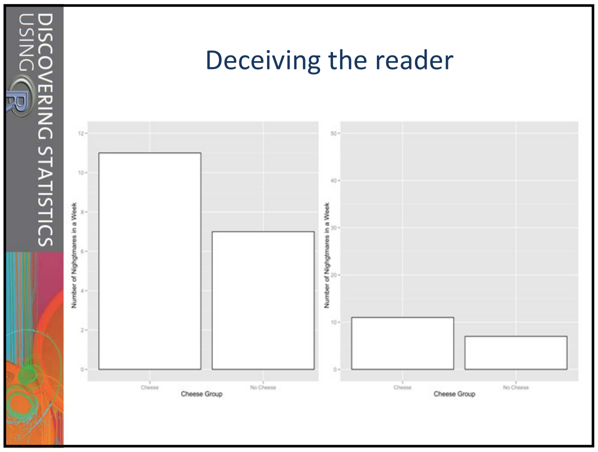
Why is this graph bad?

4



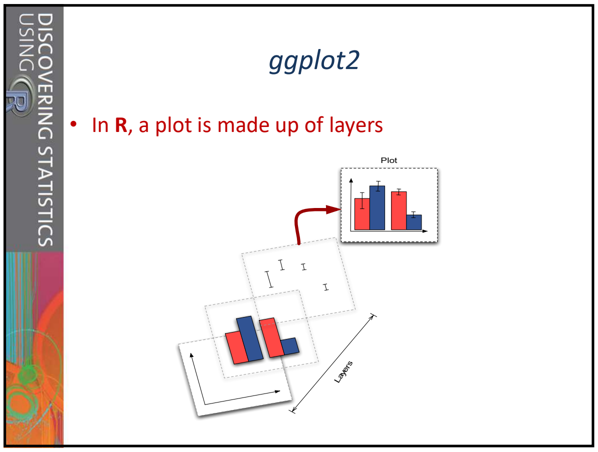
Why is this graph better?

5



Deceiving the reader

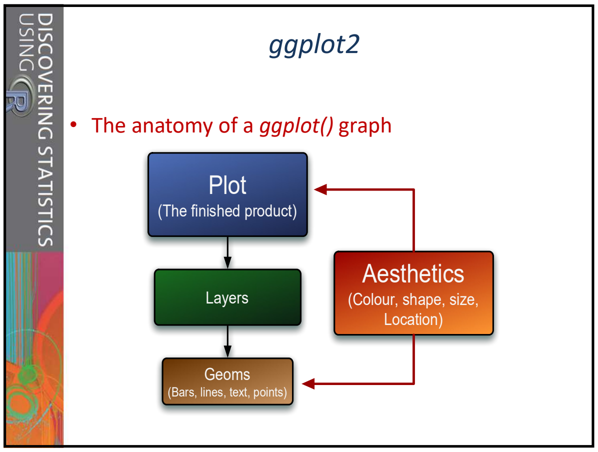
6



ggplot2

- In **R**, a plot is made up of layers

7



ggplot2

- The anatomy of a *ggplot()* graph

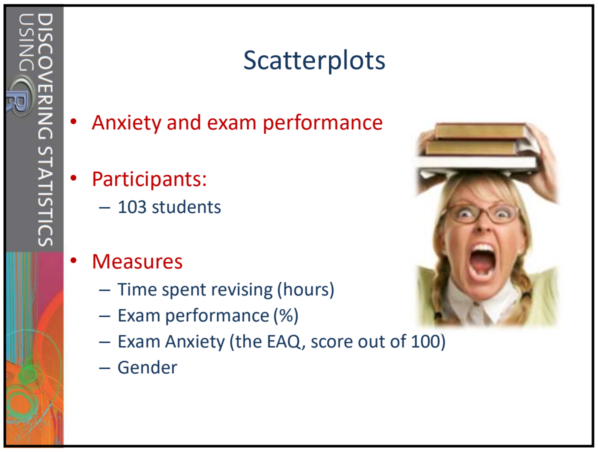
Plot
(The finished product)

Layers

Geoms
(Bars, lines, text, points)

Aesthetics
(Colour, shape, size, Location)

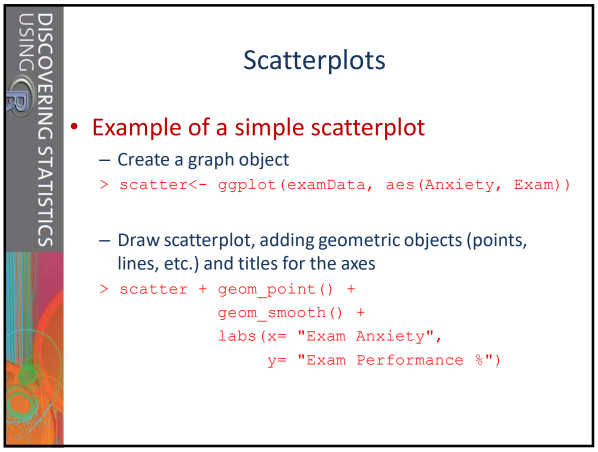
8


DISCOVERING STATISTICS USING R

Scatterplots

- Anxiety and exam performance
- Participants:
  - 103 students
- Measures
  - Time spent revising (hours)
  - Exam performance (%)
  - Exam Anxiety (the EAQ, score out of 100)
  - Gender

9


DISCOVERING STATISTICS USING R

Scatterplots

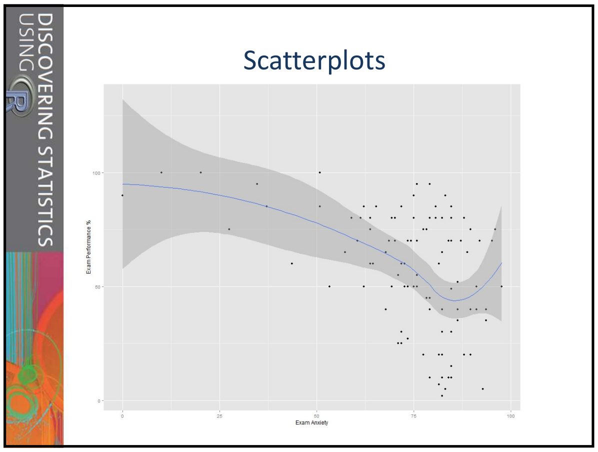
- Example of a simple scatterplot
  - Create a graph object

```
> scatter<- ggplot(examData, aes(Anxiety, Exam))
```

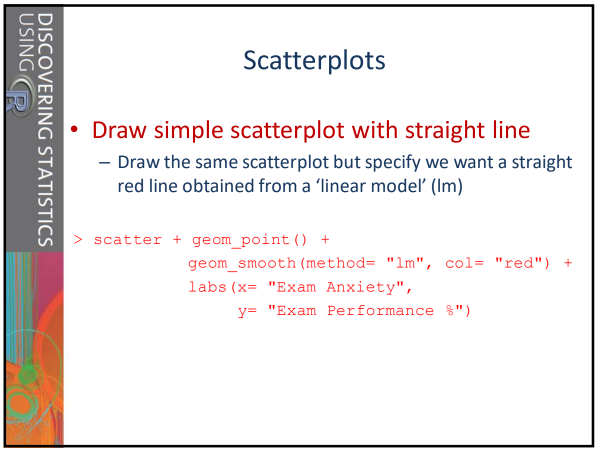
  - Draw scatterplot, adding geometric objects (points, lines, etc.) and titles for the axes

```
> scatter + geom_point() +
            geom_smooth() +
            labs(x= "Exam Anxiety",
                 y= "Exam Performance %")
```

10



11

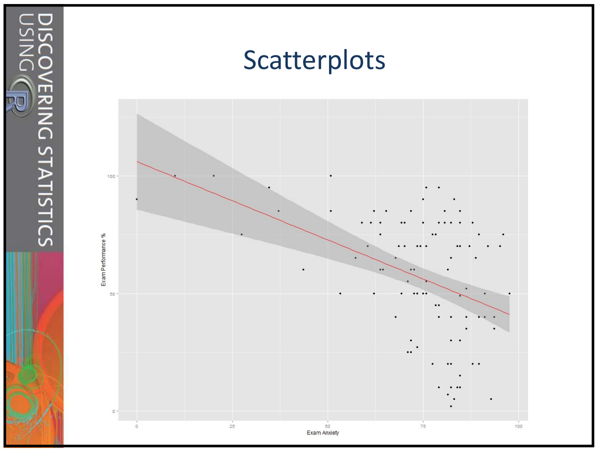

DISCOVERING STATISTICS USING R

Scatterplots

- Draw simple scatterplot with straight line
  - Draw the same scatterplot but specify we want a straight red line obtained from a ‘linear model’ (lm)

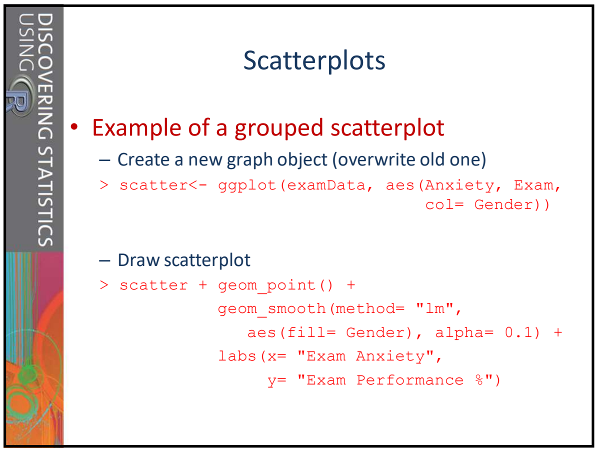
```
> scatter + geom_point() +
          geom_smooth(method= "lm", col= "red") +
          labs(x= "Exam Anxiety",
               y= "Exam Performance %")
```

12



Scatterplots

13



Scatterplots

- Example of a grouped scatterplot
  - Create a new graph object (overwrite old one)

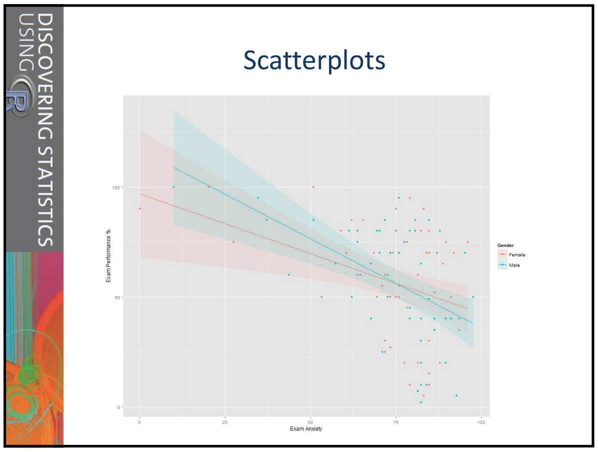
```
> scatter<- ggplot(examData, aes(Anxiety, Exam,
                                  col= Gender))
```

  - Draw scatterplot

```
> scatter + geom_point() +
            geom_smooth(method= "lm",
               aes(fill= Gender), alpha= 0.1) +
            labs(x= "Exam Anxiety",
                 y= "Exam Performance %")
```

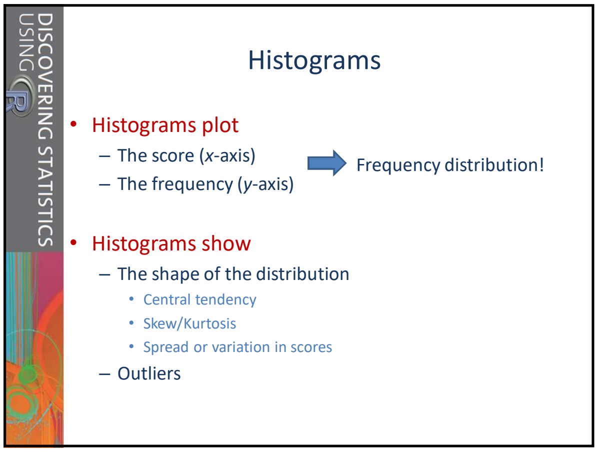
DISCOVERING STATISTICS USING R

14



Scatterplots

15



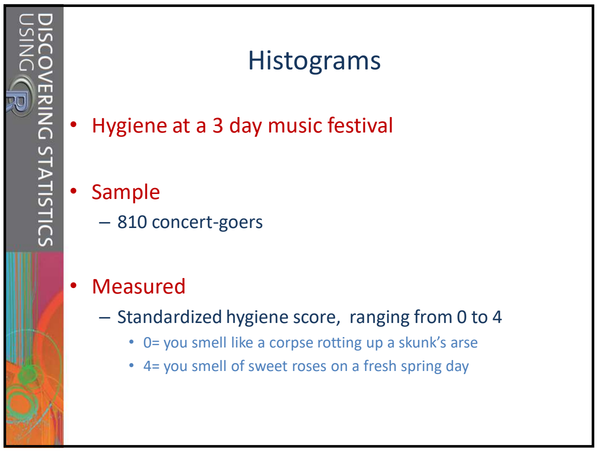
Histograms

- Histograms plot
  - The score (*x*-axis)
  - The frequency (*y*-axis)

→ Frequency distribution!

- Histograms show
  - The shape of the distribution
    - Central tendency
    - Skew/Kurtosis
    - Spread or variation in scores
  - Outliers

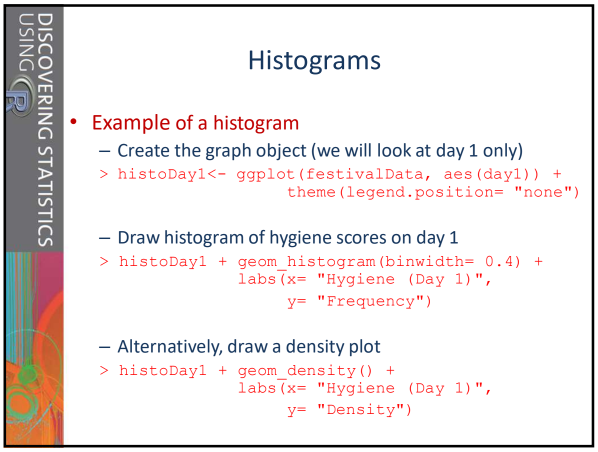
16



Histograms

- Hygiene at a 3 day music festival
- Sample
  - 810 concert-goers
- Measured
  - Standardized hygiene score, ranging from 0 to 4
    - 0= you smell like a corpse rotting up a skunk's arse
    - 4= you smell of sweet roses on a fresh spring day

17



Histograms

- Example of a histogram
  - Create the graph object (we will look at day 1 only)

```
> histoDay1<- ggplot(festivalData, aes(day1)) +
                  theme(legend.position= "none")
```

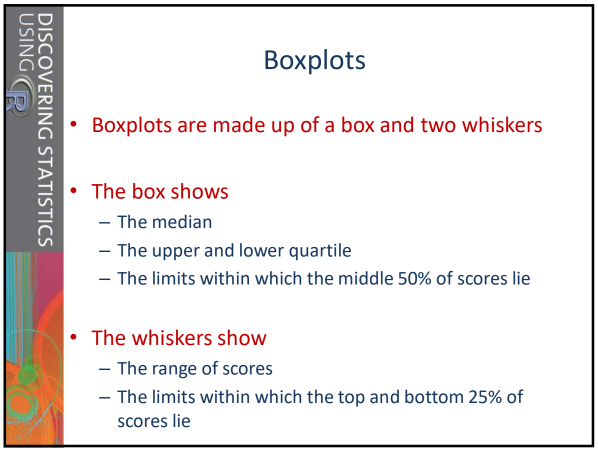
  - Draw histogram of hygiene scores on day 1

```
> histoDay1 + geom_histogram(binwidth= 0.4) +
             labs(x= "Hygiene (Day 1)",
                  y= "Frequency")
```

  - Alternatively, draw a density plot

```
> histoDay1 + geom_density() +
             labs(x= "Hygiene (Day 1)",
                  y= "Density")
```

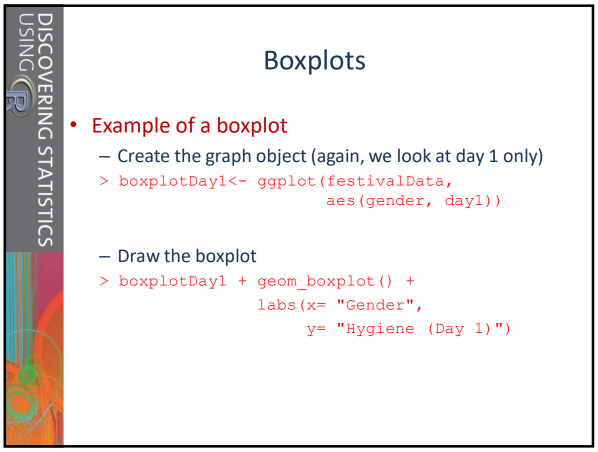
18


DISCOVERING STATISTICS USING R

Boxplots

- Boxplots are made up of a box and two whiskers
- The box shows
  - The median
  - The upper and lower quartile
  - The limits within which the middle 50% of scores lie
- The whiskers show
  - The range of scores
  - The limits within which the top and bottom 25% of scores lie

19


DISCOVERING STATISTICS USING R

Boxplots

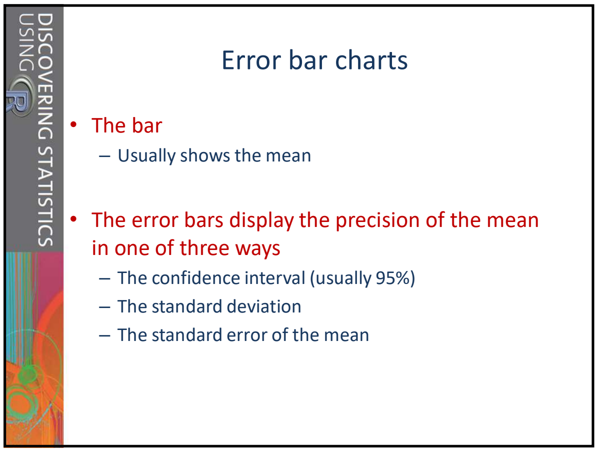
- Example of a boxplot
  - Create the graph object (again, we look at day 1 only)

```
> boxplotDay1<- ggplot(festivalData,
                       aes(gender, day1))
```

  - Draw the boxplot

```
> boxplotDay1 + geom_boxplot() +
               labs(x= "Gender",
                    y= "Hygiene (Day 1)")
```

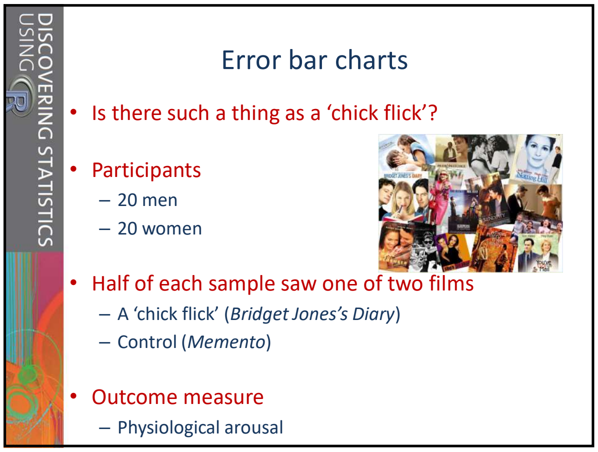
20


DISCOVERING STATISTICS USING R

Error bar charts

- The bar
  - Usually shows the mean
- The error bars display the precision of the mean in one of three ways
  - The confidence interval (usually 95%)
  - The standard deviation
  - The standard error of the mean

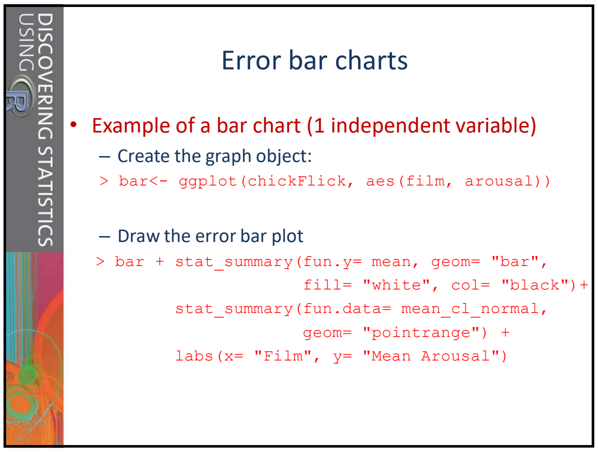
21



Error bar charts

- Is there such a thing as a ‘chick flick’?
- Participants
  - 20 men
  - 20 women
- Half of each sample saw one of two films
  - A ‘chick flick’ (*Bridget Jones’s Diary*)
  - Control (*Memento*)
- Outcome measure
  - Physiological arousal

22



Error bar charts

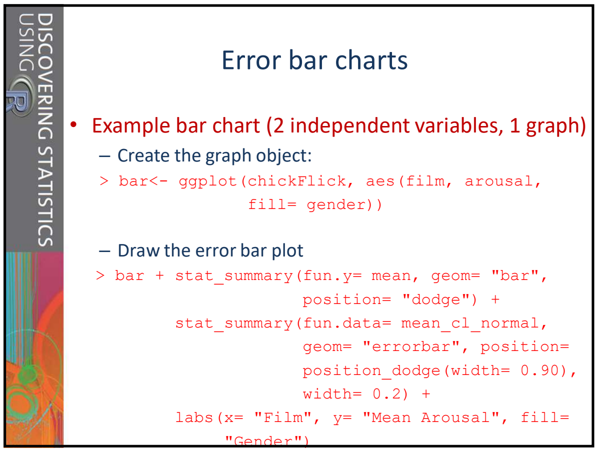
- Example of a bar chart (1 independent variable)
  - Create the graph object:

```
> bar<- ggplot(chickFlick, aes(film, arousal))
```

  - Draw the error bar plot

```
> bar + stat_summary(fun.y= mean, geom= "bar",
                   fill= "white", col= "black")+
        stat_summary(fun.data= mean_cl_normal,
                   geom= "pointrange") +
        labs(x= "Film", y= "Mean Arousal")
```

23



Error bar charts

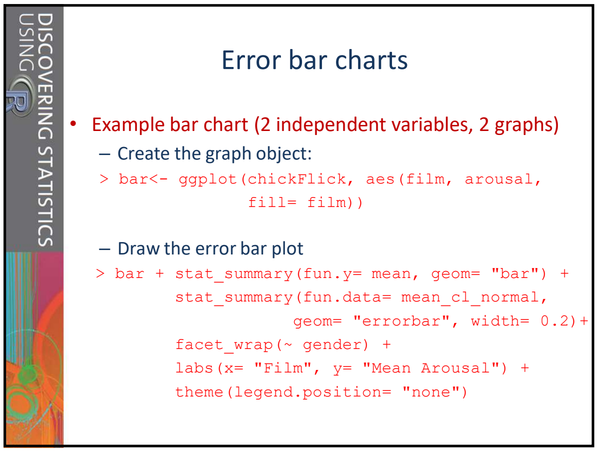
- Example bar chart (2 independent variables, 1 graph)
  - Create the graph object:

```
> bar<- ggplot(chickFlick, aes(film, arousal,
              fill= gender))
```

  - Draw the error bar plot

```
> bar + stat_summary(fun.y= mean, geom= "bar",
                   position= "dodge") +
        stat_summary(fun.data= mean_cl_normal,
                   geom= "errorbar", position=
                   position_dodge(width= 0.90),
                   width= 0.2) +
        labs(x= "Film", y= "Mean Arousal", fill=
            "Gender")
```

24


DISCOVERING STATISTICS USING R

Error bar charts

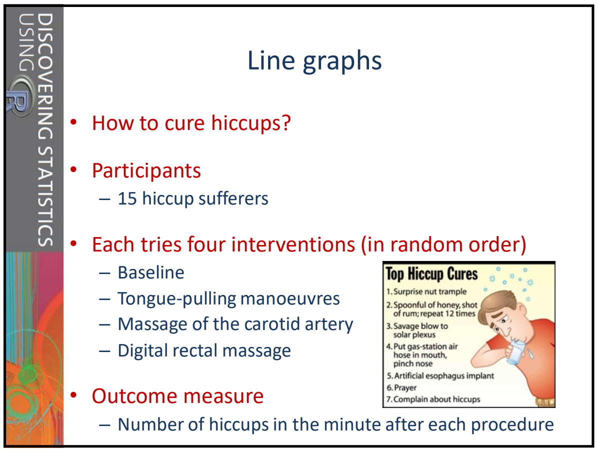
- Example bar chart (2 independent variables, 2 graphs)
  - Create the graph object:

```
> bar<- ggplot(chickFlick, aes(film, arousal,
               fill= film))
```

  - Draw the error bar plot

```
> bar + stat_summary(fun.y= mean, geom= "bar") +
        stat_summary(fun.data= mean_cl_normal,
                     geom= "errorbar", width= 0.2)+
        facet_wrap(~ gender) +
        labs(x= "Film", y= "Mean Arousal") +
        theme(legend.position= "none")
```

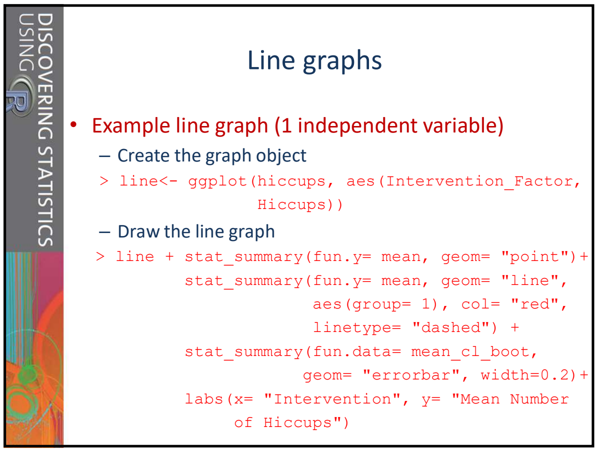
25


DISCOVERING STATISTICS USING R
Top Hiccup Cures
1. Surprise nut trample
2. Spoonful of honey, shot of rum; repeat 12 times
3. Savage blow to solar plexus
4. Put gas-station air hose in mouth, pinch nose
5. Artificial esophagus implant
6. Prayer
7. Complain about hiccups

Line graphs

- How to cure hiccups?
- Participants
  - 15 hiccup sufferers
- Each tries four interventions (in random order)
  - Baseline
  - Tongue-pulling manoeuvres
  - Massage of the carotid artery
  - Digital rectal massage
- Outcome measure
  - Number of hiccups in the minute after each procedure

26


DISCOVERING STATISTICS USING R

Line graphs

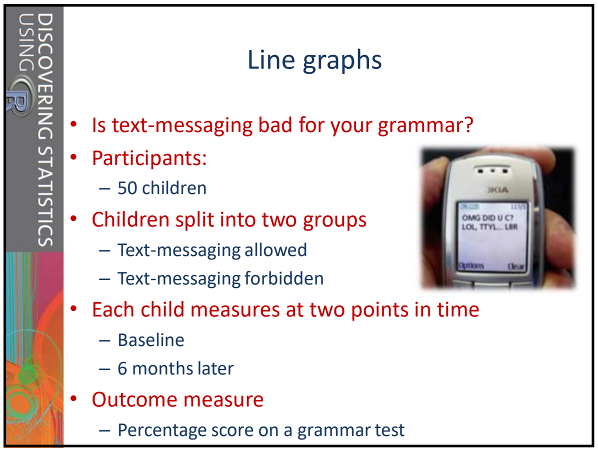
- Example line graph (1 independent variable)
  - Create the graph object

```
> line<- ggplot(hiccups, aes(Intervention_Factor,
                Hiccups))
```

  - Draw the line graph

```
> line + stat_summary(fun.y= mean, geom= "point")+
         stat_summary(fun.y= mean, geom= "line",
                      aes(group= 1), col= "red",
                      linetype= "dashed") +
         stat_summary(fun.data= mean_cl_boot,
                     geom= "errorbar", width=0.2)+
         labs(x= "Intervention", y= "Mean Number
              of Hiccups")
```

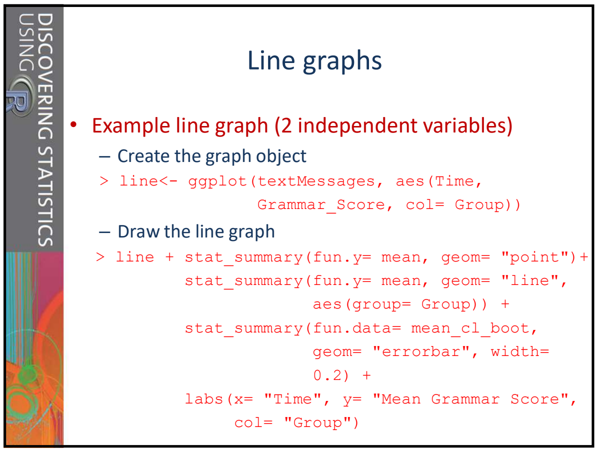
27


DISCOVERING STATISTICS USING R
OMG DID U C? LOL, TTYL... L8R

Line graphs

- Is text-messaging bad for your grammar?
- Participants:
  - 50 children
- Children split into two groups
  - Text-messaging allowed
  - Text-messaging forbidden
- Each child measures at two points in time
  - Baseline
  - 6 months later
- Outcome measure
  - Percentage score on a grammar test

28


DISCOVERING STATISTICS USING R

Line graphs

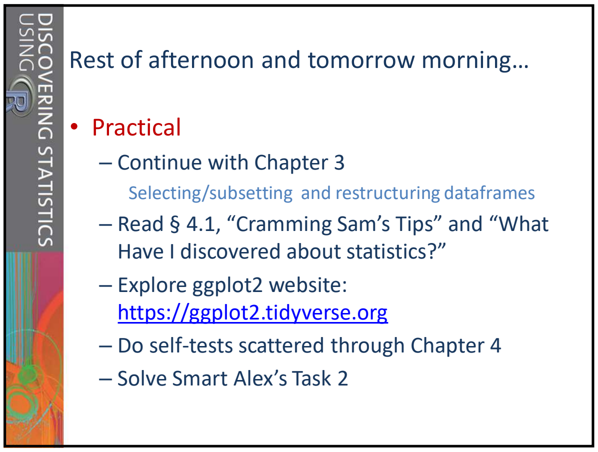
- Example line graph (2 independent variables)
  - Create the graph object

```
> line<- ggplot(textMessages, aes(Time,
                 Grammar_Score, col= Group))
```

  - Draw the line graph

```
> line + stat_summary(fun.y= mean, geom= "point")+
         stat_summary(fun.y= mean, geom= "line",
                      aes(group= Group)) +
         stat_summary(fun.data= mean_cl_boot,
                      geom= "errorbar", width=
                      0.2) +
         labs(x= "Time", y= "Mean Grammar Score",
              col= "Group")
```

29

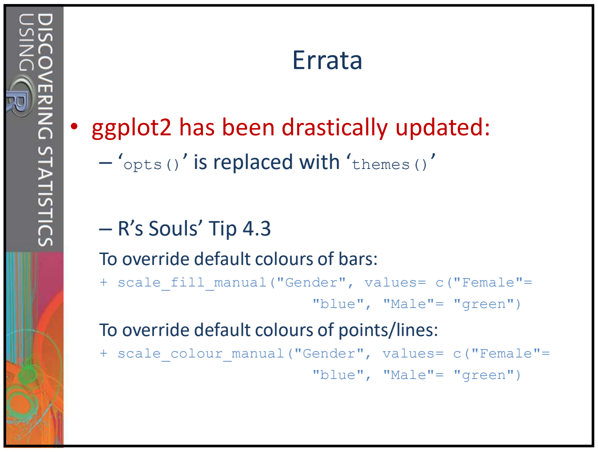

DISCOVERING STATISTICS USING R

Rest of afternoon and tomorrow morning…

- Practical
  - Continue with Chapter 3

    Selecting/subsetting and restructuring dataframes
  - Read § 4.1, "Cramming Sam's Tips" and "What Have I discovered about statistics?"
  - Explore ggplot2 website: https://ggplot2.tidyverse.org
  - Do self-tests scattered through Chapter 4
  - Solve Smart Alex's Task 2

30



DISCOVERING STATISTICS USING R

Errata

- ggplot2 has been drastically updated:
  - ‘`opts()`’ is replaced with ‘`themes()`’
  - R’s Souls’ Tip 4.3

To override default colours of bars:

```
+ scale_fill_manual("Gender", values= c("Female"=
                       "blue", "Male"= "green")
```

To override default colours of points/lines:

```
+ scale_colour_manual("Gender", values= c("Female"=
                       "blue", "Male"= "green")
```

31