February 23

- Finish up a few topics from Chapter 2
- Introduce Data Visualization (Chapter 4)
- Demonstration of ggplot2
- Practical tomorrow
  – some end-of-chapter exercises from Chapter 4
- Homework
  – no written homework this week
  – read Chapters 4 & 5

Going beyond the data: *z*-scores

- *z*-scores
  - Expresses a data point in terms of how many standard deviations it is away from the mean.
  - The distribution of *z*-scores has a mean of 0 and *SD* = 1.

$$z = \frac{X - \bar{X}}{s}$$

Z-scores: Examples

- If you had a mean of 5 and a standard deviation of 2, what would the z-score be of:
  - a datapoint with a value of 8?
  - a datapoint with a value of 1?
  - a datapoint with a value of 6?

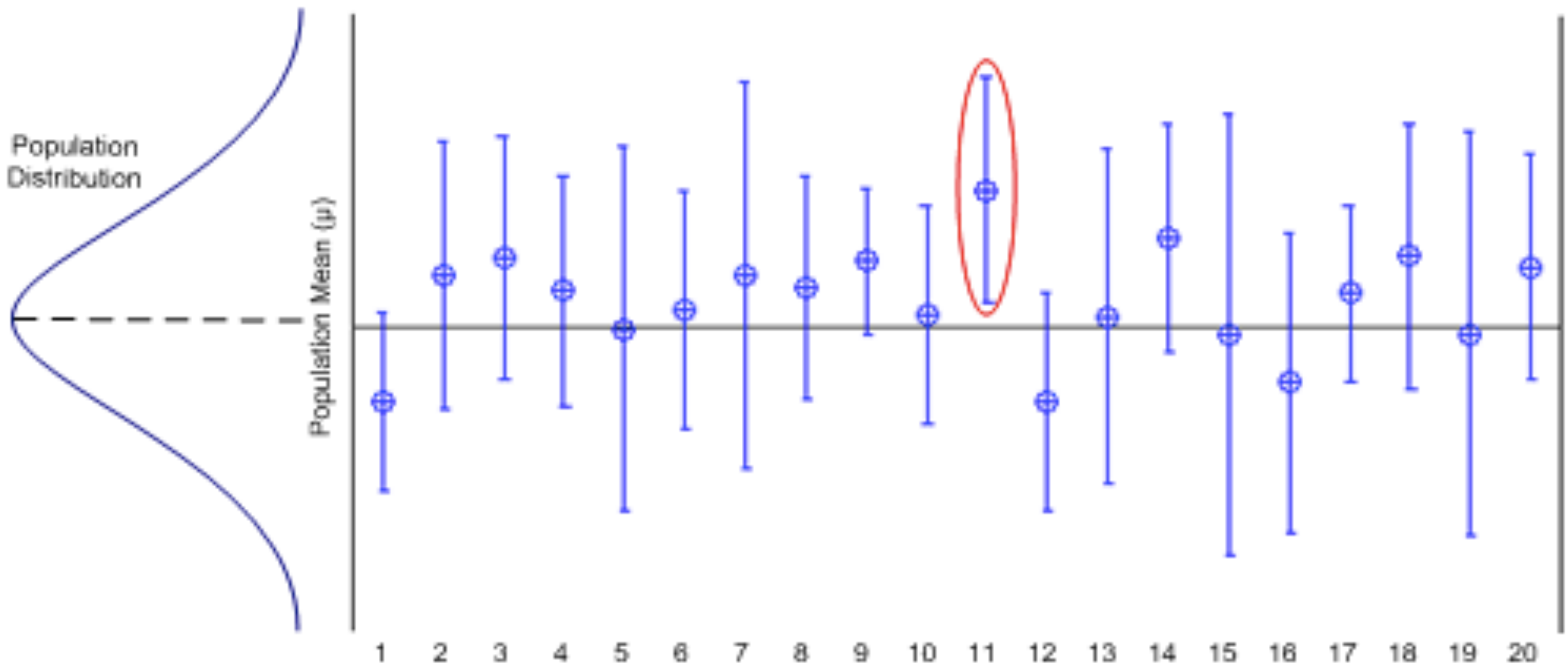
Properties of *z*-scores

- 1.96 cuts off the top 2.5% of the distribution.
- −1.96 cuts off the bottom 2.5% of the distribution.
- As such, 95% of *z*-scores lie between −1.96 and 1.96.
- 99% of *z*-scores lie between −2.58 and 2.58.
- 99.9% of them lie between −3.29 and 3.29.

Confidence Intervals

- True mean (not directly observable)
  - Happiness score of 15
- Sample mean
  - Happiness score of 17
- Interval estimate
  - 12 to 22 (contains true value)
  - 16 to 18 (misses true value)
  - CIs constructed such that 95% of time they contain the true mean.

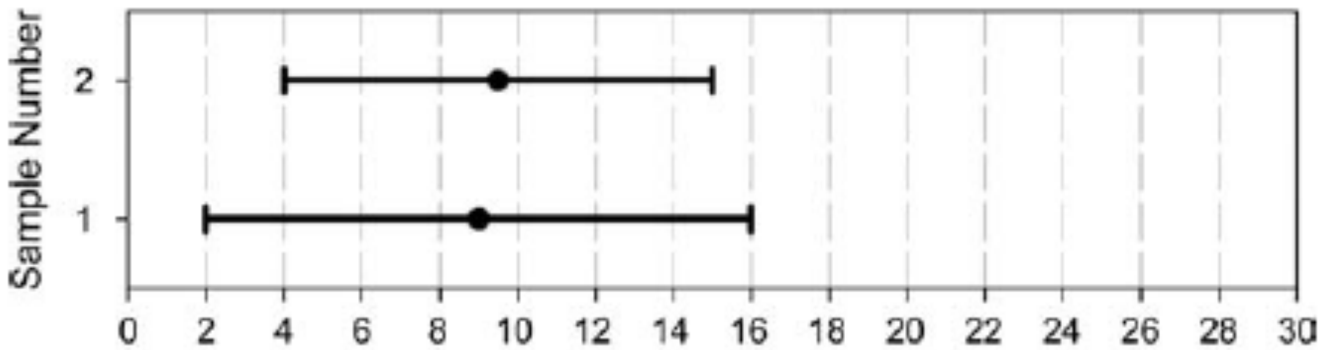
Confidence Intervals for 20 samples

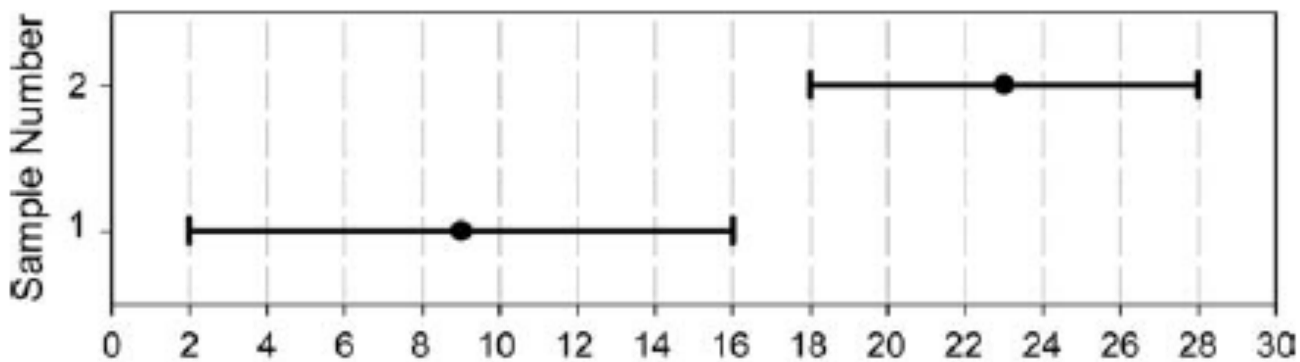


Calculating Confidence Intervals

- Remember critical values from z-scores…
  – What do -1.96 and 1.96 represent?
- Boundaries of confidence intervals
  1. lower = sample mean – 1.96*Standard Error
  2. upper = sample mean + 1.96*Standard Error
- What about for 99% confidence intervals?

Happiness Scores





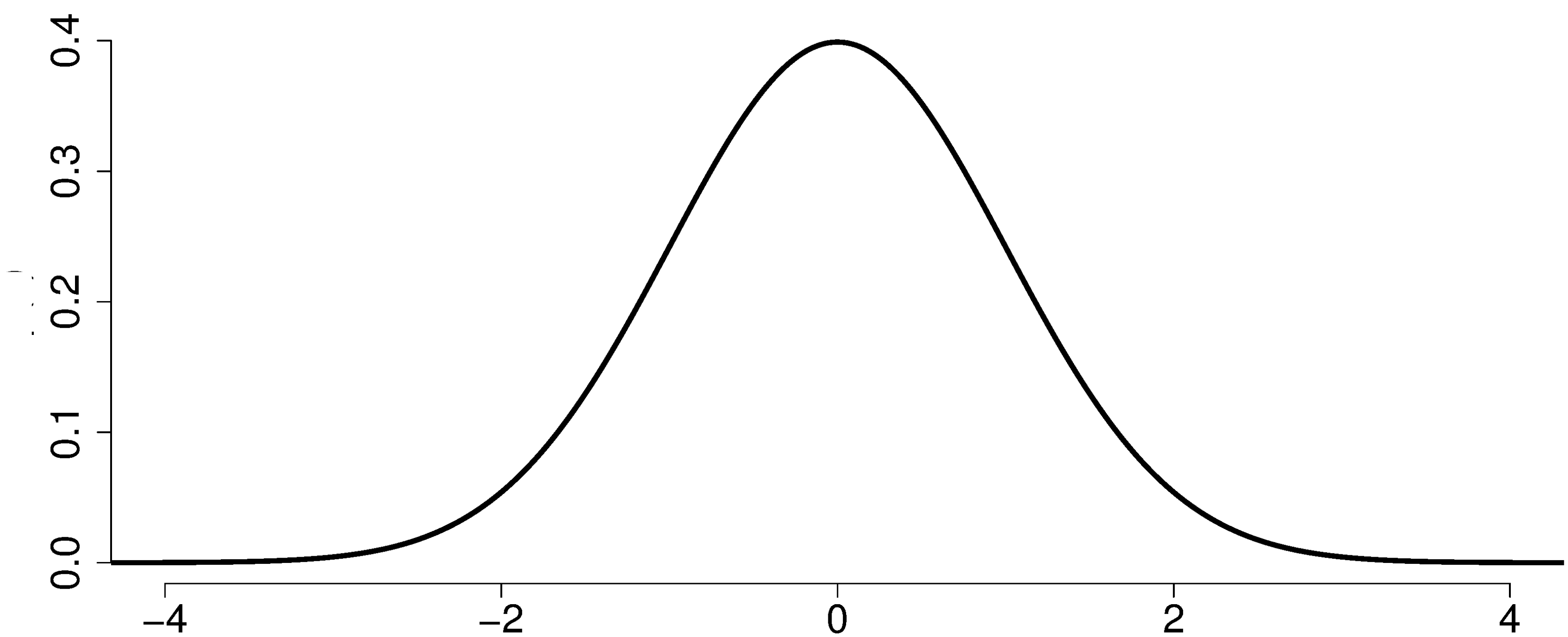
Test Statistics

- Measures that compare systematic to unsystematic variation
  - t, F, chi-square,

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

  - The probability of particular values is known

Example: T distribution



Test Statistics, cont.

- allow us to say whether the fit of our models to the data is *significant*
  - significant: when a test score cuts off less than the top 5% of probability (critical value of the test statistic)
    - This ratio of explained to unexplained variation is highly unlikely
    - our model would be unlikely to fit this well if there was no experimental effect

Types of Hypotheses

- Null hypothesis, H_0
  - There is no effect.
  - E.g. *Big Brother* contestants and members of the public will not differ in their scores on personality disorder questionnaires
- The alternative hypothesis, H_1
  - Aka the experimental hypothesis
  - E.g. *Big Brother* contestants will score higher on personality disorder questionnaires than members of the public

Directional and Non-directional hypotheses

- Big Brother contestants will score higher on a personality disorder questionnaire than the general public
- Big Brother contestants will score lower on a personality disorder questionnaire than the general public
- Big Brother contestants will score differently on a personality disorder questionnaire than the general public

One- and Two-Tailed Tests

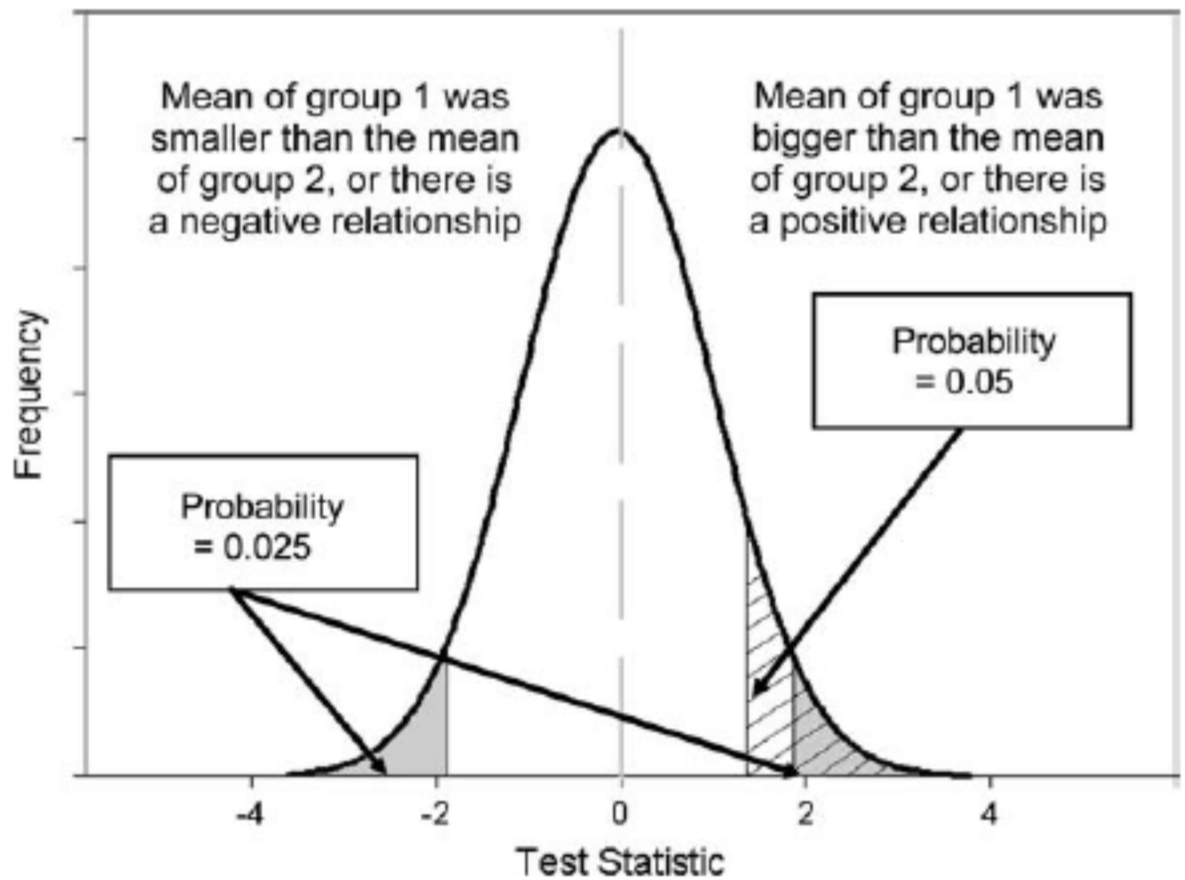


FIGURE 2.10 Diagram to show the difference between one- and two-tailed tests

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

Type I and Type II Errors

- Type I error
  - occurs when we believe that there is a genuine effect in our population when, in fact, there isn't.
- Type II error
  - occurs when we believe that there is no effect in the population when, in reality, there is.

Effect Sizes

- An effect size is a standardized measure of the size of an experimental effect:
  - Standardized = comparable across studies with different units of measurement
  - Allows people to objectively evaluate the size of observed effect.
  - Very small and unimportant effects can be significant if huge sample size

Effect Size Measures (Cohen's d, Pearson's r…)

- *r* = .1, *d* = .2 (small effect):
  - the effect explains 1% of the total variance.
- *r* = .3, *d* = .5 (medium effect):
  - the effect accounts for 9% of the total variance.
- *r* = .5, *d* = .8 (large effect):
  - the effect accounts for 25% of the variance.
- Beware of these 'canned' effect sizes though:
  - The size of effect should be placed within the research context.

Exploring Data with Graphs

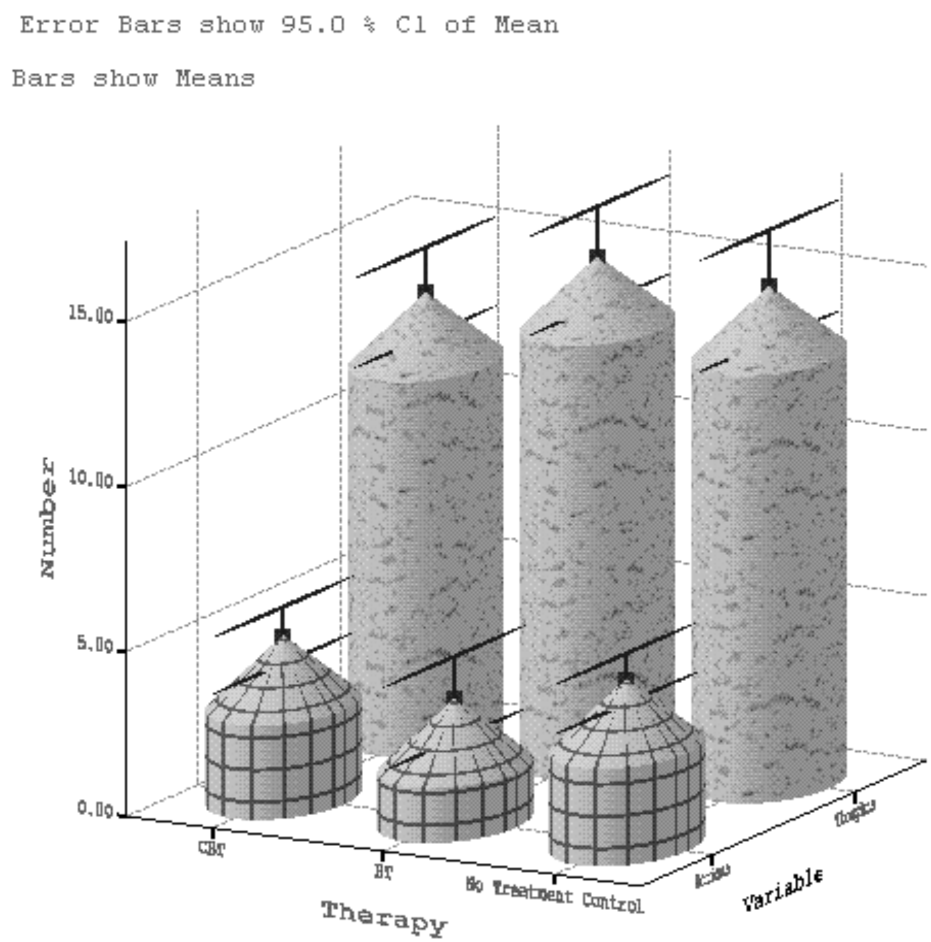
Aims

- How to present data clearly
- Introduce *ggplot2*
- Graphs
  - Scatterplots
  - Histograms
  - Boxplots
  - Error bar charts
  - Line graphs

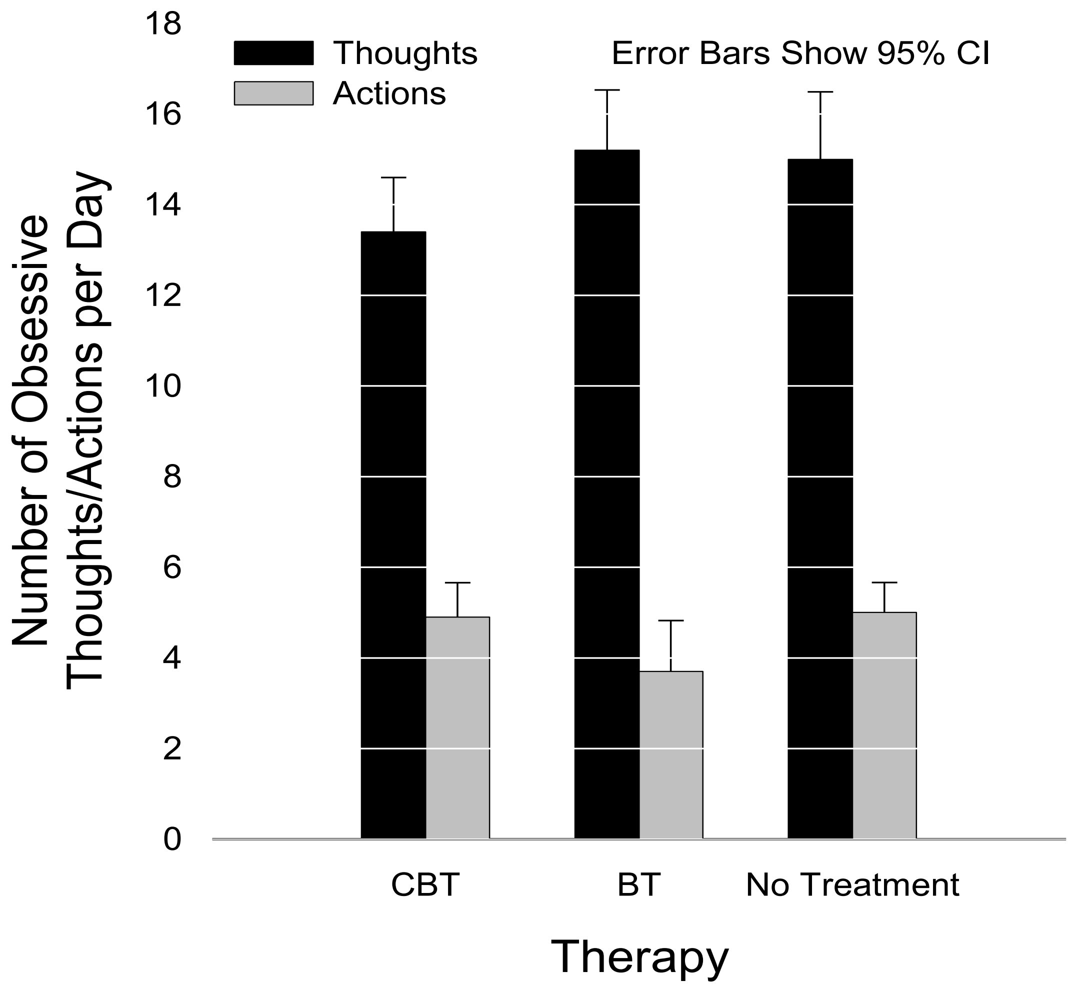
The Art of Presenting Data

- Graphs should (Tufte, 2001):
  - Show the data.
  - Induce the reader to think about the data being presented (rather than some other aspect of the graph).
  - Avoid distorting the data.
  - Present many numbers with minimum ink.
  - Make large data sets (assuming you have one) coherent.
  - Encourage the reader to compare different pieces of data.

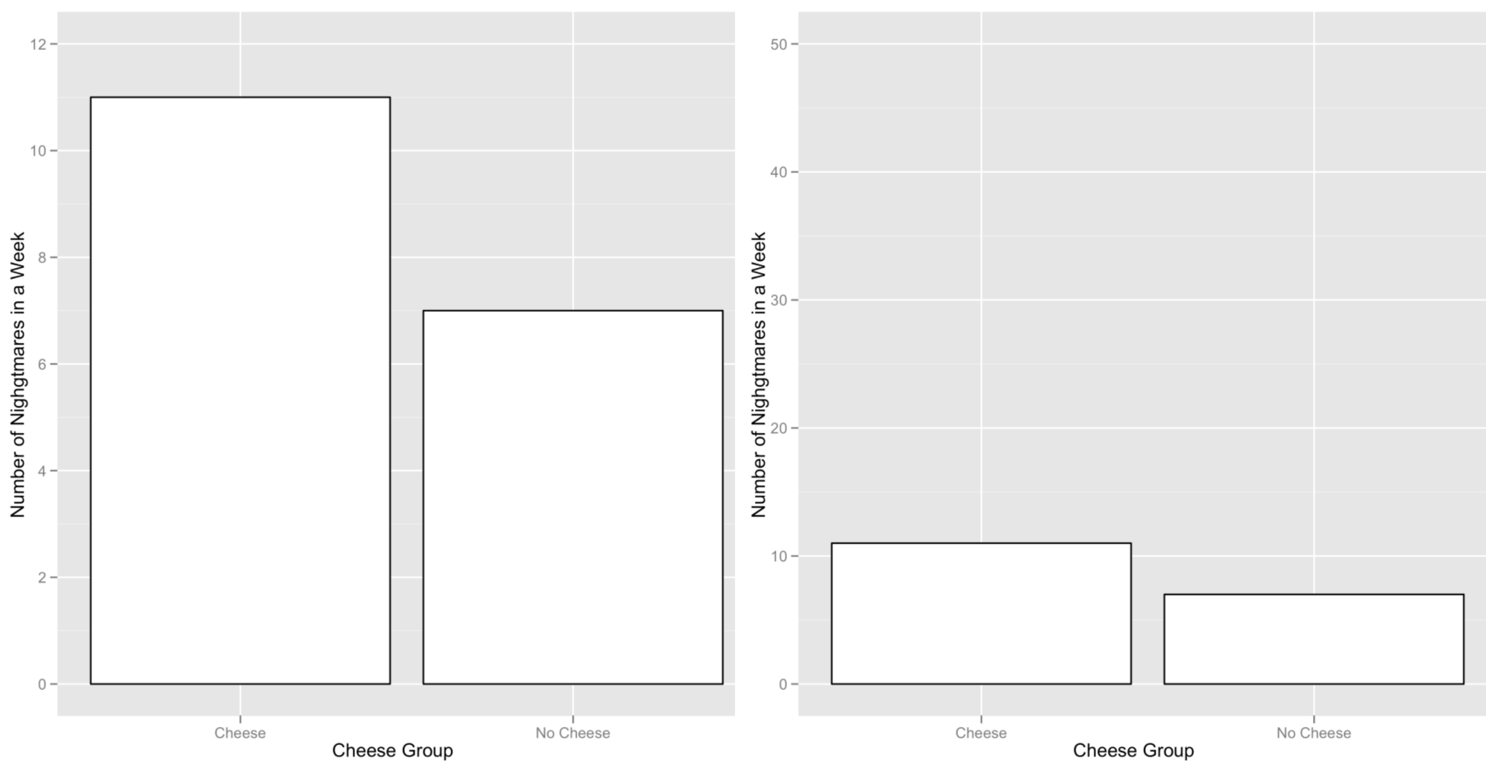
Why Is This Graph Bad?



Why Is This Graph Better?



Deceiving the Reader

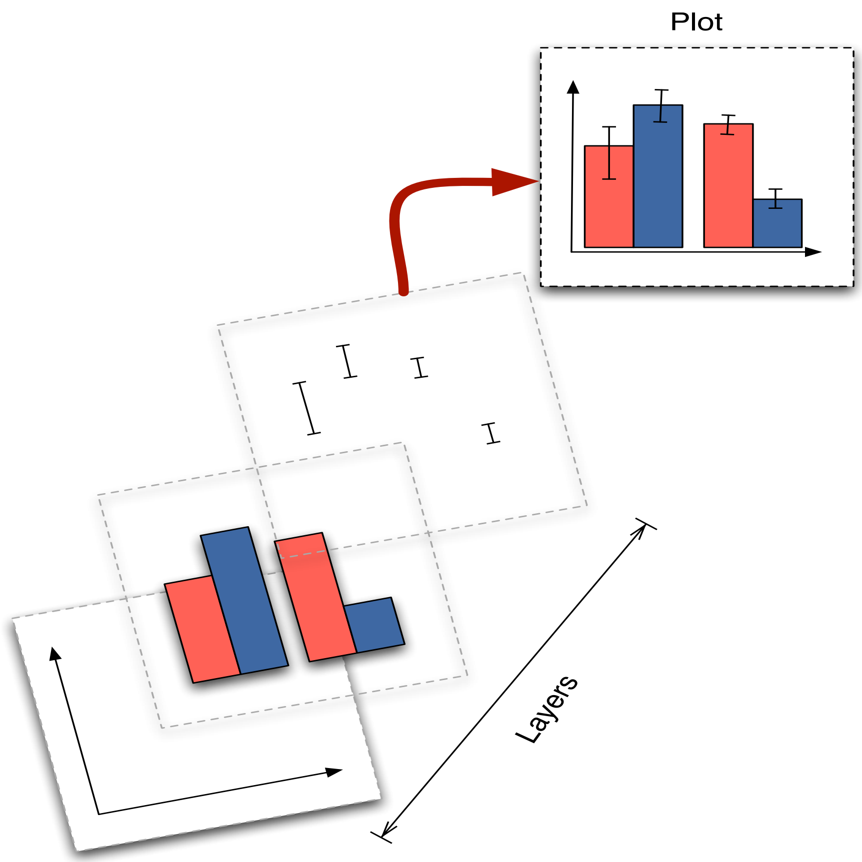


Two graphs about cheese

Plotting graphs in RStudio

- R comes with built-in functions for graphs
  - plot(), hist(), barplot(), boxplot()
  - good for quick plotting; less intuitive/flexible for more complex plotting
- We will be using ggplot2
  - powerful, versatile package for creating full range of high-quality graphs
  - Common interface
  - Bit of a learning curve

ggplot2



In *ggplot2* a plot is made up of layers.

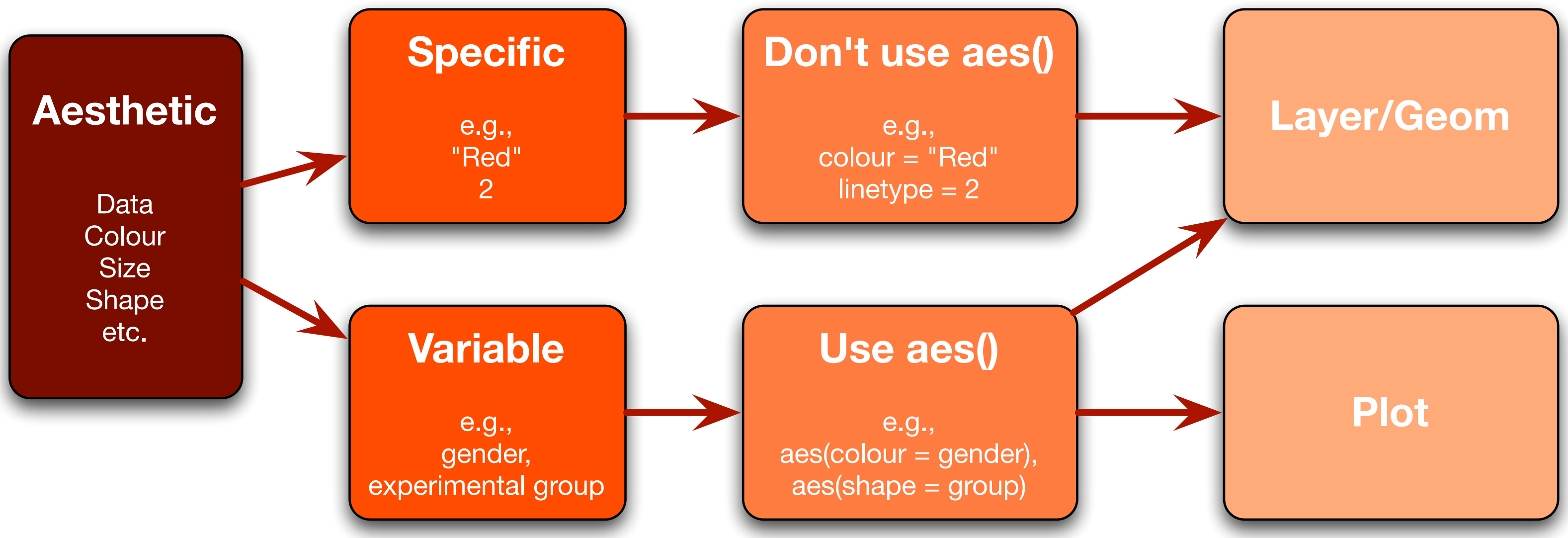
Examples of geoms (pp 124-25)

- geom_bar()
- geom_point()
- geom_line()
- geom_histogram()
- geom_text()
- geom_errorbar()
- …

Aesthetic properties

- Defined for graph as whole OR for individual geoms
  - cascade down; lower-level override
- Required and optional for different geoms

GEOM	REQUIRED	OPTIONAL
geom_point()	x: variable to plot on x-axis y: variable to plot on y-axis	shape color size fill alpha (transpar.)



Specifying aesthetics in *ggplot2*

see page 126 for difference aesthetic options

“Stats”

- Functions that are part of ggplot2 the perform particular statistical operations
  - bin data, compute quartiles, estimate density, compute central tendency measures, etc.
- Used automatically by geoms
- Used by us when we want to override defaults or provide additional information in plots