MEDB 5501, Module04

2024-09-10

Topics to be covered

- What you will learn
  - Scatterplots and correlations
  - Boxplots and effect sizes
  - R code for scatterplots and boxplots
  - Bar plots for categorical outcomes
  - R code for bar plots

Assessing relationships (bivariate statistics), 1

- Between two continuous variables
  - Scatterplot
  - Correlation

Speaker notes

The previous modules talked about how to examine individual variables. I used the term univariate statistics. After you understand how each individual variable behaves, you need to start examining how variables relate to one another. I will focus on graphical approaches, but also some numeric measures.

Recall that continuous variables have a large number of possible values, potentially any value in some interval. Categorical variables have a small number of possible values with each value corresponding to a particular label.

If you are visualizing the relationship between two continuous variable, use a scatterplot. The best numeric summary is a correlation coefficient.

Assessing relationships (bivariate statistics), 2

- Between a categorical and a continuous variable
  - Boxplot
  - Effect size

Speaker notes

If you are visualizing the relationship between a categorical and a continuous variable, use a boxplot. A common numeric summary is the effect size. I do not like effect sizes, but they are used so commonly that I have to teach them.

Assessing relationships (bivariate statistics), 3

- Between two categorical variables
  - Bar plots
  - Many numeric measures (covered later)

Speaker notes

If you are visualizing the relationship between two categorical variables, use a bar plot of the probabilities. There are many numeric measures, and I will defer discussion of these until a later module.

Scatterplot

- Assess relationship between two continuous variables
  - Outcome on y-axis
  - Exposure/intervention on x-axis

Speaker notes

The scatterplot is a simple and easy way to examine the relationship between two independent variables. By tradition (and for no other good reason), the outcome variable is plotted on the y-axis (the vertical axis) and the exposure or treatment variable is plotted on the x-axis (the horizontal axis). If you conceive of the relationship as cause and effect, the cause variable goes on the x-axis Sometimes it is not clear which is which. You have to use your best judgement. I will usually provide some guidance in the programming assignments.

Patterns on a scatterplot

- Strong, weak, or no relationship
- Direction of relationship
- Nonlinear patterns
  - Diminishing returns
  - Exponential acceleration
  - Hormesis

Speaker notes

Scatterplots require you to make a subjective interpretation. Is the relationship strong or weak? Is there any relationship at all?

What is the direction of the relationship? Does an increase in exposure lead to an increase in the outcome? That’s a positive relationship. If an increase in exposure leads to a decrease in the outcome, that’s a negative relationship.

Look for nonlinear patterns. These can take many forms, but three common ones are diminishing returns, exponential acceleration, or hormesis. I’ll show examples of these in a few minutes.

Covariance

- $Cov(X,Y) = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$
  - $X_i > \bar{X}$ and $Y_i > \bar{Y}$; (+) × (+) = +
  - $X_i < \bar{X}$ and $Y_i < \bar{Y}$; (-) × (-) = +
  - $X_i > \bar{X}$ and $Y_i < \bar{Y}$; (+) × (-) = -
  - $X_i < \bar{X}$ and $Y_i > \bar{Y}$; (-) × (+) = -

Speaker notes

Your book introduces the term covariance first before defining correlation. The covariance is a product involving the individual X and Y values compared to their respective means.

There are four possibilities. An individual X_i and Y_i values could both be above average. This makes the terms (X_i-Xbar) and (Y_i - Ybar) both positive. When you multiply two positive numbers together, you get a positive result.

A second possibility is that both individual values are below average. This makes the terms (X_i-Xbar) and (Y_i - Ybar) both negative. When you multiply two negative numbers you get a positive result.

If most of the data follows this pattern, above average values of X associated with above average values of Y and below average values of X associated with below average values of Y, then the covariance will be positive.

You might have the opposite pattern occur. Above average values of X are mostly associated with below average values of Y and vice versa. In this case, the covariance will be negative.

The correlation coefficient, 1

- $r = Corr(X, Y) = \frac{Cov(X,Y)}{S_X S_y}$
  - r is always between -1 and +1.
  - r is a unitless quantity

Speaker notes

The correlation coefficient between X and Y is computed as the covariance between X and Y divided by the standard deviation of X and the standard deviation of Y.

This creates a unitless quantity. While the covariance will change when you convert measurements from grams to kilograms, the correlation will not.

In general, I am not in favor of unitless quantities. They may allow you to compare different outcomes, but to understand the practical implications, you need to use a measure with units. I'll address this a bit later with the discussion about effect sizes.

The correlation coefficient, 2

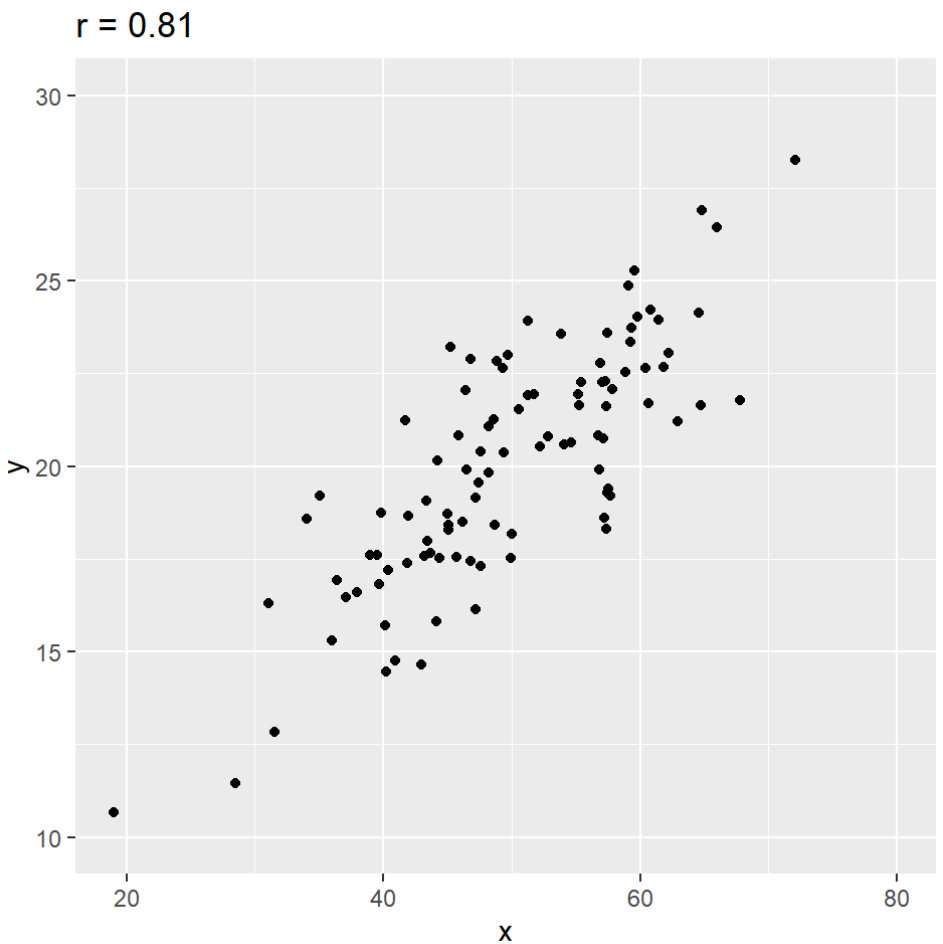
- Informal interpretation
  - +0.7 < r < +1.0 is a strong positive relationship
  - +0.3 < r < +0.7 is a weak positive relationship
  - -0.3 < r < +0.3 is little or no relationship
  - -0.7 < r < -0.3 is a weak negative relationship
  - -1.0 < r < -0.7 is a strong negative relationship

Speaker notes

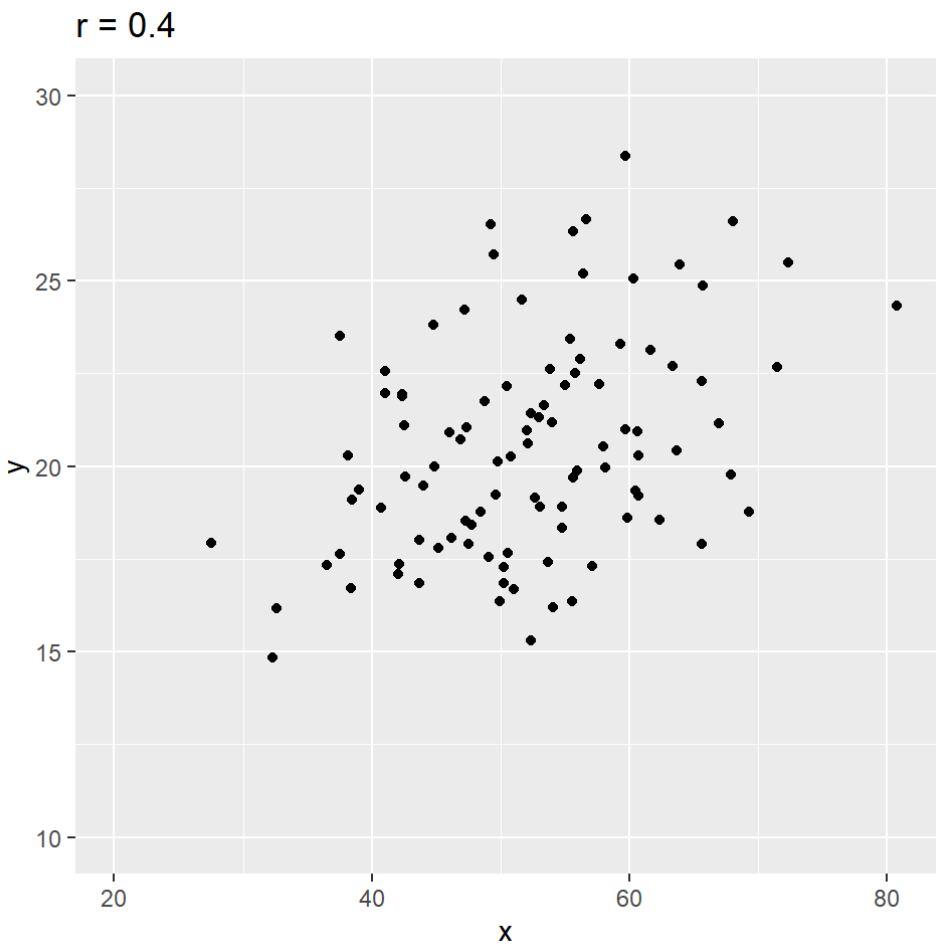
The closer the correlation is to +1, the stronger the positive relationship. The closer the correlation is to -1, the stronger the negative relationship. A correlation close to 0 implies little or no relationship.

There is some debate about the dividing lines between strong and weak relationships, but a dividing line around +/-0.6 or 0.7 is common. Anything closer to zero than +/-0.2 or 0.3 is considered evidence of little or no relationship.

An example of a strong positive relationship

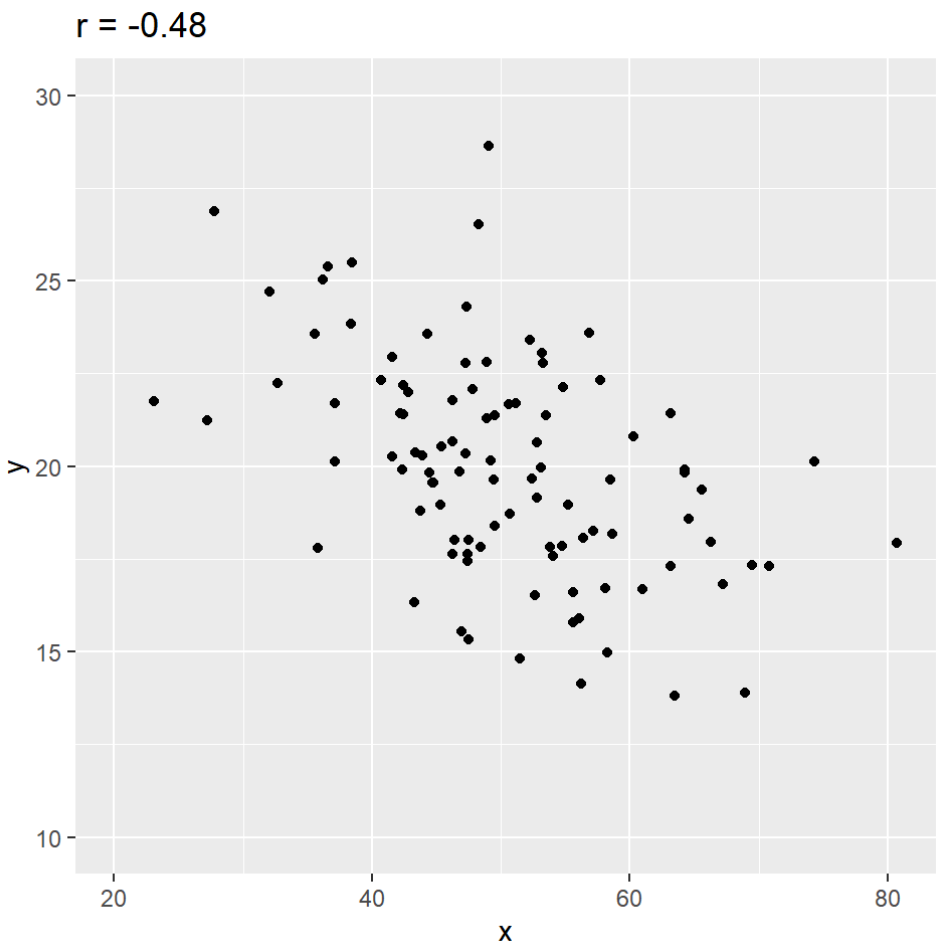


An example of a weak positive relationship

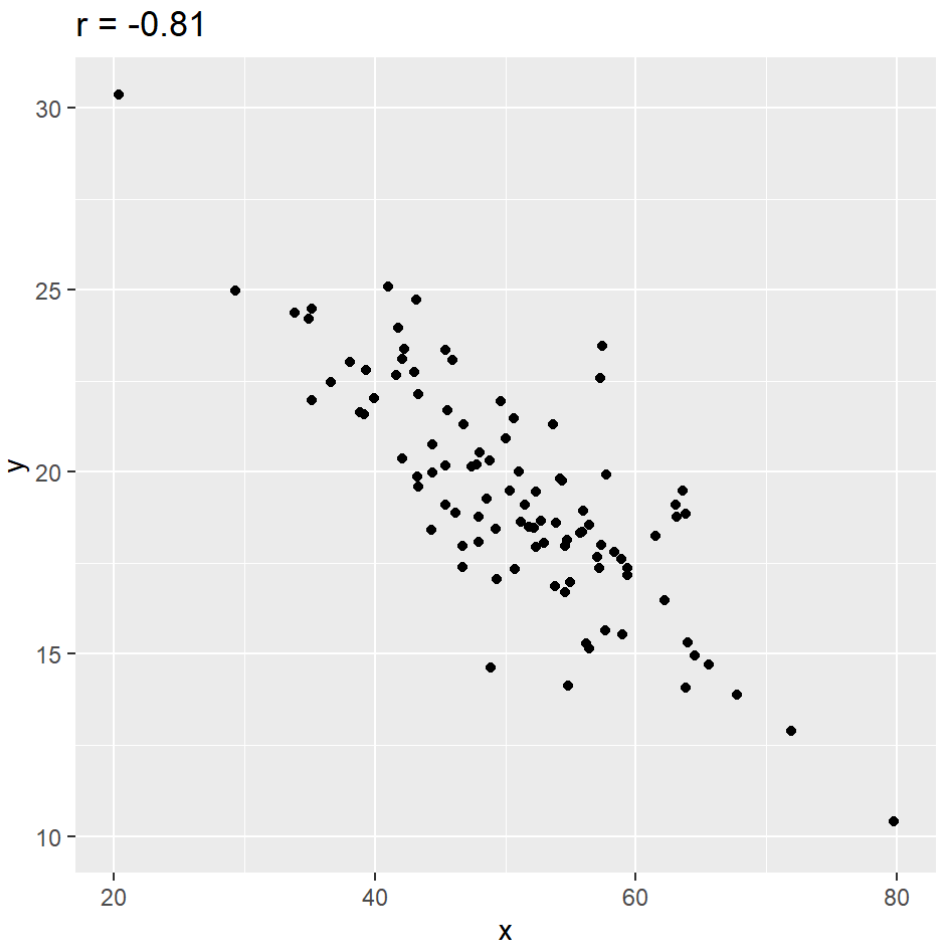


An example of a little or no relationship

An example of a weak negative relationship



An example of a strong negative relationship



Break #1

- What you have learned
  - Scatterplots and correlations
- What’s coming next
  - Boxplots and effect sizes

Boxplots

- Five number summary
  - Min, 25%, 50%, 75%, Max
- Assess relationship between categorical and continuous outcome.
  - One boxplot for each category
- Look for shift from one boxplot to another
  - Also changes in variation

Speaker notes

Earlier, I described the boxplot as a method for assessing whether a variable was roughly normally distributed. You can also use them to see how a continuous variable changes for different levels of a categorical variable.

Look for a shift from one boxplot to another. Do you see larger values for all five numbers in one category compared to those numbers in the other category?

Effect size, 1

- $ES = \frac{\bar{X}_1 - \bar{X}_2}{S}$
  - Different choices for S
  - S_1, S_2, or S_p
- Also known as standardized mean difference (SMD)

Speaker notes

The effect size is a measure of how much difference there is between the average value for one category compared to the average value for a second category. It is divided by the standard deviation to make it a unitless quantity.

There are different choices for which standard deviation to use. If the two groups have roughly the same amount of variation, it does not matter. Sometimes, though, one group is more variable than the other. In this case, you have to choose the standard deviation of the first group, the standard deviation of the second group, or a compromise, the pooled standard deviation, which will be partway between the two different standard deviations. I will talk about the pooled standard deviation in a later lecture. If one of the two categories could be considered a comparison or control group, you might use the standard deviation for that category.

Effect size, 1

- Interpretation (controversial!)
  - Small: ES = 0.2
    - Height difference between 15 and 16 year old girls
  - Medium: ES = 0.5
    - Height difference between 14 and 18 year old girls
  - Large: ES = 0.8
    - Height difference between 13 and 18 year old girls

Speaker notes

The interpretation of the effect size was developed by Jacob Cohen in his famous book of 1988, Statistical Power Analysis for the Behavioral Sciences.

To try to visualize this, Jacob Cohen talked about height differences between girls of different ages.

Uses of the effect size

- Input statistic for systematic overviews/meta-analyses
- Intermediate endpoint in power and sample size calculations
- Direct estimate of power and sample size (NO!!)
- Gauging the practical significance of a research study (NO!!)

Speaker notes

The effect size provides a useful way to compare results from individual studies in a systematic overview or meta-analysis. The effect size allows you to compare apples and oranges, so to speak, though you still need to be careful. I've said many times that combining apples and oranges is often okay, but you shouldn't be combining apples and onions.

The effect size is an important intermediate endpoint in estimates of power and sample size. You will see a lot about power and sample size in later weeks, but I wanted to address this point now.

Some researchers will use the effect size as a direct estimate of power and sample size and this is not a good idea. I will try to explain why in the next slide.

Another practice that I do not like is when research will use an effect size to gauge whether the results of a completed research study have practical significance.

Some researchers will also use the effect size to gauge the practical significance of a research study.

Criticisms of the effect size

- Sample size calculations need three things
  - Research hypothesis
  - Variation of the outcome measure
  - Minimum clinically important difference (MCID)
    - Unitless quantities have no clinical meaning
- Small, medium, large vary across disciplines
- Provide cover for arbitrary sample size choices
  - You must consider variation and MCID separately

Speaker notes

A proper assessment of sample size requires three things (sometimes four, but most of the time, just three). The third item that you need is the minimum clinically important difference. What is clinically important? It is something that requires clinical judgement. I am not a clinician, but I have seen examples where this is completely overlooked. A researcher will say that any difference is clinically important. This is never true, and you need to think carefully about this.

The effect size has no clinical meaning. I tell a joke about a large store that puts up a banner saying “Big weekend sale! All prices reduced by half a standard deviation.”

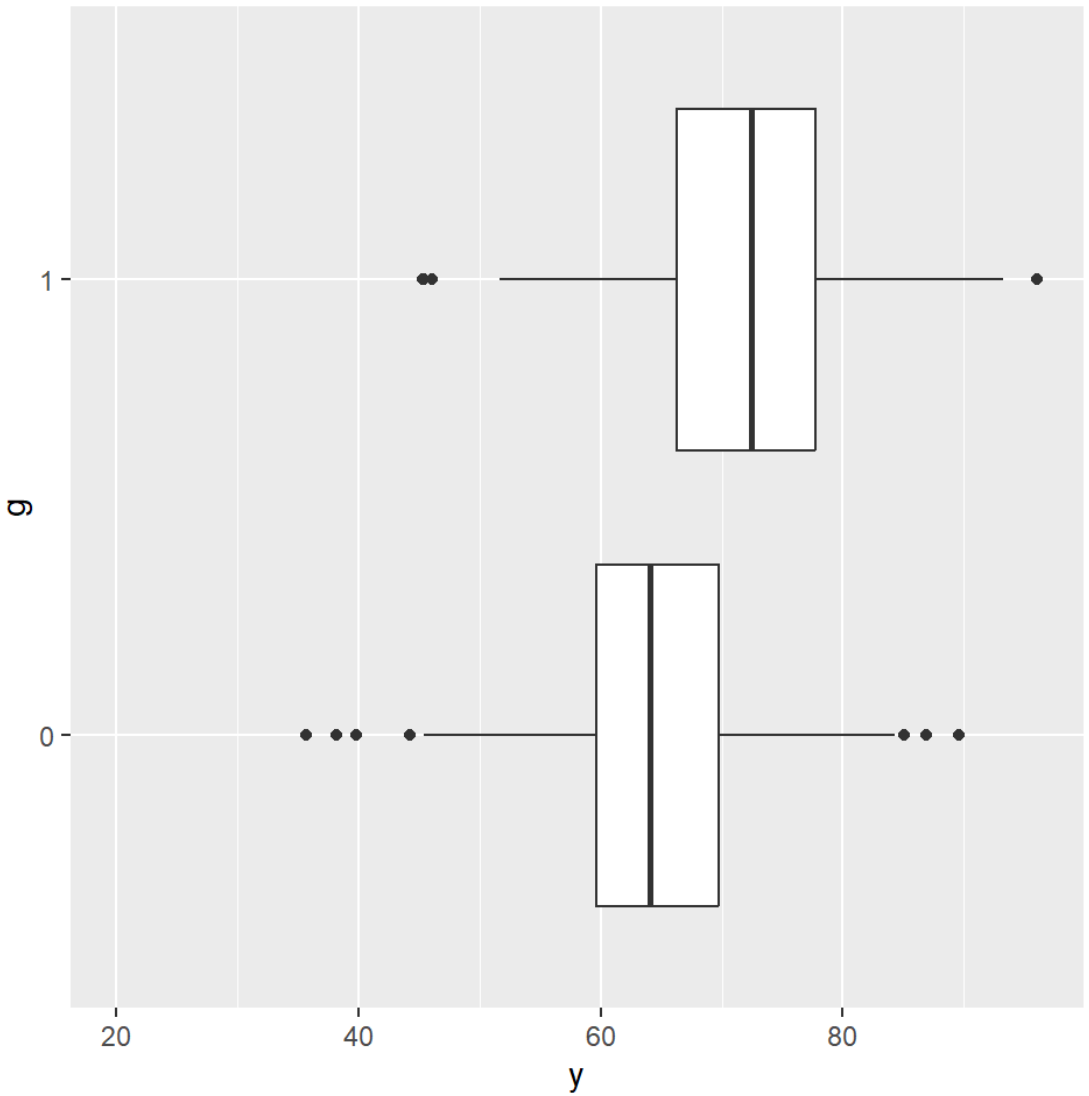
The concept of small, medium, and large effect sizes is also controversial. These were developed by Jacob Cohen after reviewing a bunch of psychology studies. The review is now several decades old, and it didn’t really consider studies in other fields like business, education, or health care.

The effect size also combines the variation and the MCID into a single number. But when you are planning a research study, there are things that you can do that influence the variation, such as better equipment and training. There are things that you can do to influence the MCID, such as using a measure that is more responsive to real changes in health.

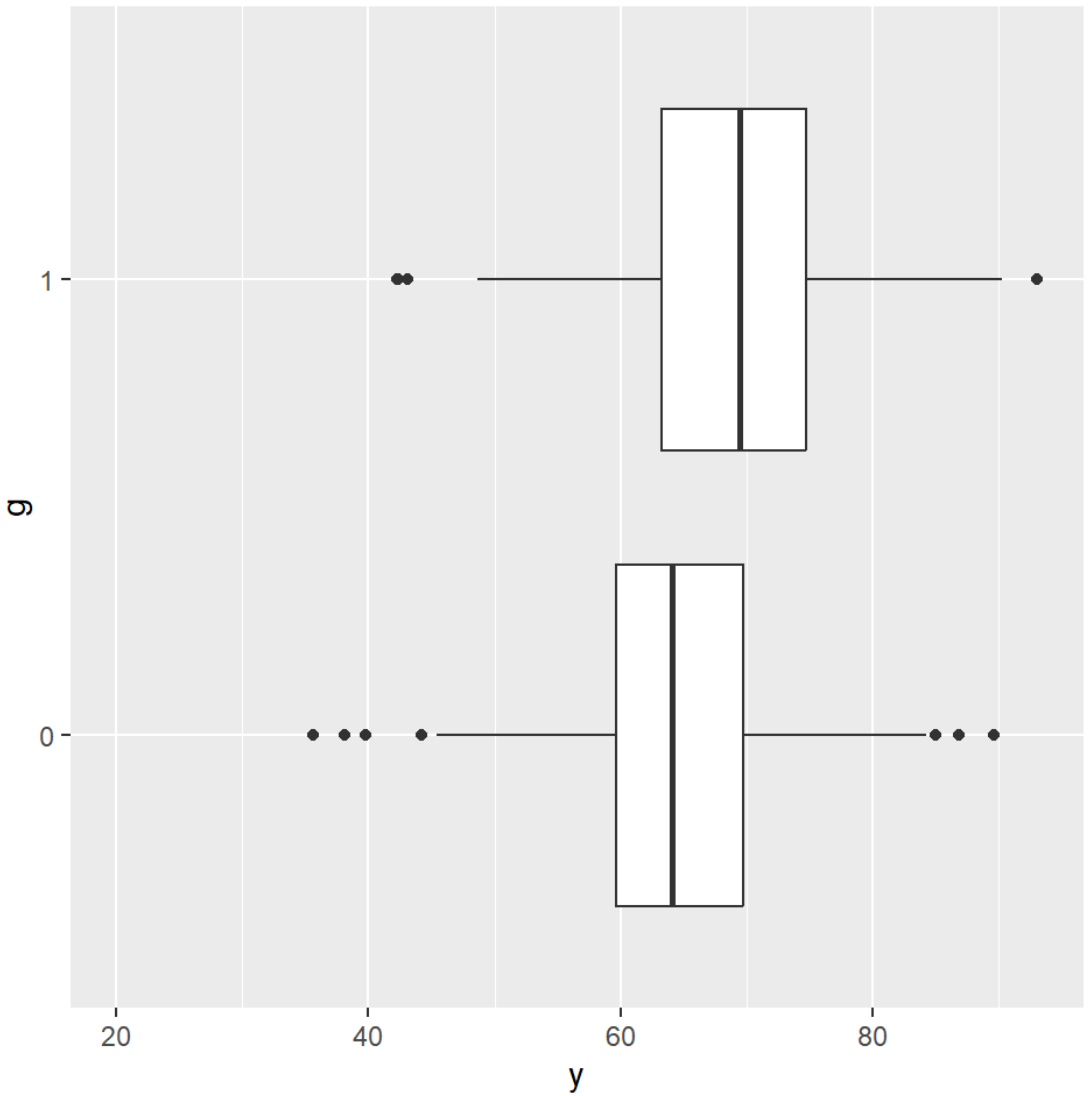
If you combine the variation and MCID into a single number and use that number to drive your power and sample size calculations, you lose the opportunity to explore important aspects of the research study.

I am teaching the effect size because a majority of researchers like it. I find myself in the minority here. I am not alone and there are a lot of very smart people who have sharply criticized the use of effect sizes, except as an intermediate calculation or as the unit of analysis for systematic overviews.

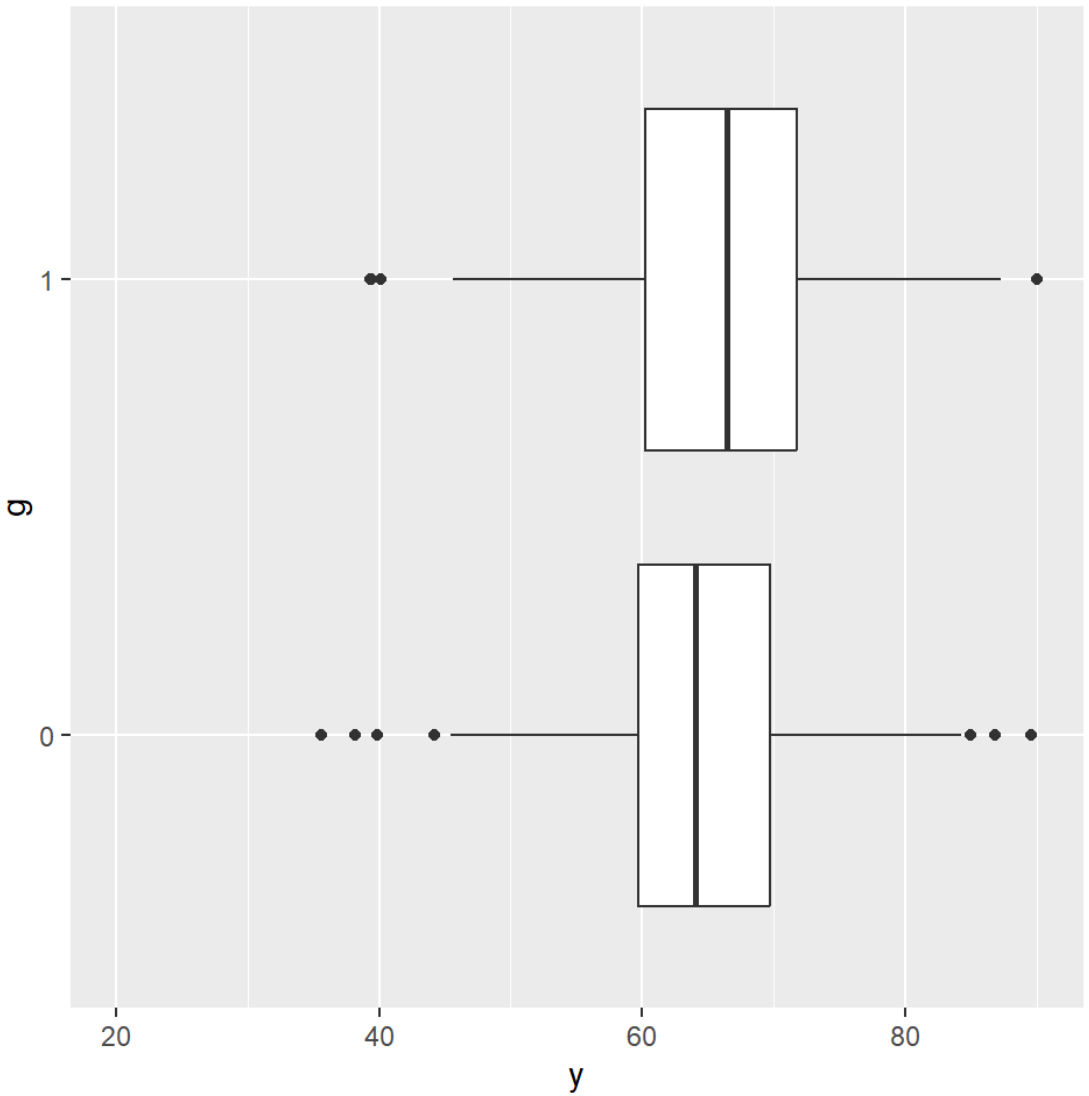
Example of a large effect size



Example of a medium effect size



Example of a small effect size



Break #2

- What you have learned
  - Boxplots and effect sizes
- What's coming next
  - R code for scatterplots and boxplots

Data dictionary for fev, 1

```
---
data_dictionary: fev (.csv, sas7bdat, .sav, .txt)

copyright: >
  The author of the jse article holds the copyright, but does not list
  conditions under which it can be used. Individual use for educational
  purposes is probably permitted under the Fair Use provisions of
  U.S. Copyright laws.

description: >
  Forced Expiratory Volume (FEV) in children. The data was collected
  in Boston in the 1970s.

additional_description:
  https://jse.amstat.org/v13n2/datasets.kahn.html
```

Data dictionary for fev, 2

```
download_url:
  https://www.amstat.org/publications/jse/datasets/fev.dat.txt

format:
  csv: comma delimited
  sas7bdat: proprietary (SAS)
  sav: proprietary (SPSS)
  txt: fixed width


varnames:
  not included


missing_value_code:
  not needed


size:
  rows: 654
  columns: 5
```

Speaker notes

Data dictionary for fev, 3

```
vars:
  age:
    scale: ratio
    range: positive integer
    unit: years

  fev:
    label: Forced Expiratory Volume
    scale: ratio
    range: positive real
    unit: liters
```

Speaker notes

This is a small dataset with eight rows and three columns.

Data dictionary for fev, 4

```
  ht:
    label: Height
    scale: positive real
    unit: inches
  sex:
    value:
      F: Female
      M: Male
  smoke:
    value:
      N: Nonsmoker
      Y: Smoker
---
```

Speaker notes

The variables are measurements before and after a major overhaul of the air conditioning system. The units are colonies per cubic foot of air. A pump pushes a certain volume of air through a filter and then bacterial colonies are allowed to grow on that filter.

simon-5501-04-fev.qmd, 1

```
---
title: "Analysis of relationships in pulmonary data"
format:
  html:
    embed-resources: true
editor: source
---

This program assesses the relationships among variables in a study of
pulmonary function in children. There is a [data dictionary][dd] that provides
more details about the data. The program was written by Steve Simon on 2024-
09-07 and is placed in the public domain.

[dd]: https://github.com/pmean/datasets/blob/master/fev.yaml
```

simon-5501-04-fev.qmd, 2

````
## Libraries

The tidyverse library is the only one you need for  this program.

```{r setup}
#| message: false
#| warning: false
library(tidyverse)
```
````

simon-5501-04-fev.qmd, 3

````
## List variable names

Since the variable names are not listed in the data file itself, you need to
list them here.

```{r names}
pulmonary_names <- c(
    "age",
    "fev",
    "ht",
    "sex",
    "smoke")
```
````

simon-5501-04-fev.qmd, 4

````
## Reading the data

Here is the code to read the data and show a glimpse.

```{r read}
pulmonary <- read_csv(
  file="../data/fev.csv",
  col_names=pulmonary_names,
  col_types="nnncc")
glimpse(pulmonary)
```
````

simon-5501-04-fev.qmd, 5

````
## Calculate mean, quartiles, range for fev

```{r descriptive-fev}
summary(pulmonary$fev)
sd(pulmonary$fev)
```

The mean fev is 2.6 liters and the standard deviation is 0.84 liters. The fev values range from 0.8 to 5.8. I am not an expert on pulmonary function, but these values appear to be reasonable.
````

simon-5501-04-fev.qmd, 6

````
## Calculate mean, quartiles, range for age

```{r descriptive-age}
summary(pulmonary$age)
sd(pulmonary$age)
```

The mean age is 9.9 years. The youngest subject is 3 years old and the oldest is 19. This is consistent with a pediatric population.
````

simon-5501-04-fev.qmd, 7

````
## Calculate counts for smoke

```{r descriptive-smoke}
pulmonary |>
  count(smoke) |>
  mutate(total=sum(n)) |>
  mutate(pct=round(100*n/total))
```

Almost all of the subjects (90% or 589 out of 654) were non-smokers.
````

simon-5501-04-fev.qmd, 8

````
## Plot age versus fev

```{r plot-age-fev}
pulmonary |>
  ggplot(aes(age, fev)) +
    geom_point() +
    geom_smooth() +
    xlab("Age (years)") +
    ylab("Forced Expiratory Volume (Liters)") +
    ggtitle("Plot drawn by Steve Simon on 2024-09-07")
```
````

simon-5501-04-fev.qmd, 9

````
## Correlation between age and fev

```{r correlation}
cor(pulmonary$age, pulmonary$fev)
```

The correlation, 0.75, and the plot both show a strong positive association between age and fev.
````

simon-5501-04-fev.qmd, 10

````
## Plot smoke versus fev

```{r plot-smoke-fev}
pulmonary |>
  ggplot(aes(smoke, fev)) +
    geom_boxplot() +
    xlab("Smoker (Yes/No)") +
    ylab("Forced Expiratory Volume (Liters)") +
    ggtitle("Plot drawn by Steve Simon on 2024-09-07")
```

The fev values are larger for smokers versus non-smokers. This is the opposite
direction from what we expected.
````

simon-5501-04-fev.qmd, 11

```
## Means and standard deviations for smokers and non-smokers.

```{r mean-fev-by-smoke}
pulmonary |>
  group_by(smoke) |>
  summarize(
    mean_fev=mean(fev),
    sd_fev=sd(fev))
```

The average fev values is 3.1 for smokers and much smaller, 2.6, for non-smokers. This is also opposite from what we expected. The standard deviations, 0.82 and 0.86, are roughly equal.
```

Break #3

- What you have learned
  - R code for scatterplots and boxplots
- What's coming next
  - Bar plots for categorical outcomes

Relationship between two categorical outcomes

- Simple counts and percentages
  - Remember that counting requires a precise definition
- Arrange your counts in a rectangle
- Calculate percents
  - Row percents add up to 100% within a row
  - Column percents add up to 100% within a column
  - Cell percents add up to 100% across the entire table.

Example of a rectangular grid of counts

income * mood Crosstabulation

Count

		mood		
		Happy	Miserable	Total
income	Rich	30	10	40
	Poor	90	70	160
Total		120	80	200

Rule #1: Never display more than one type of percent.

Table of SEX by RESP			
SEX(Sex of Patient)	RESP(Response of Patient)		
Frequency Percent Row Pct Col Pct	0	1	Total
F	17 17.00 30.36 77.27	39 39.00 69.64 50.00	56 56.00
M	5 5.00 11.36 22.73	39 39.00 88.64 50.00	44 44.00
Total	22 22.00	78 78.00	100 100.0

Speaker notes

Statistical software like SAS can produce counts, row percents, column percents, cell percents, expected counts, residuals, and/or cell contribution to chi-squared values. At one time or another you might want to use each of these statistics, but never all at one time. Two or more numbers in a table causes confusion and makes your tables harder to interpret.

Present a single summary statistic in the table if at all possible. If you need to display two summary statistics (for example, both counts and row percentages), then place the counts in one table and the row percentages in a different table. If you have to fit them in the same table, place the two numbers side by side with the less important number appearing second and in parentheses For example, 54% (257).

This table, an example of how bad the default option is for SAS, is taken from

Joseph J. Guido. Guido's Guide to PROC FREQ – A Tutorial for Beginners Using the SAS® System. Northeast SAS Users Group Conference, 2007. Available at https://www.lexjansen.com/nesug/nesug07/ff/ff07.pdf

Rule #2: Row percentages are usually best.

- Divide by row totals
- Percentages add up to 100% within each row

Speaker notes

Row percentages are the percentages you compute by dividing each count by the row total. Row percentages place the comparison between two numbers within a single column, so that one number is directly beneath the number you want to compare it to. This is usually better than column percents, where the numbers you want to compare are side by side. If you find that column percentages make more sense. Consider swapping the rows and columns.

If you find that cell percentages make the most sense, consider creating composite categories that combine the row and column categories. Cell percentages are the percentages that you get when you divide each cell count by the overall total. When cell percents are interesting, it usually means that you are interested in the four distinct categories in your two by two table. For example, you are interested in seeing what fraction of job candidates are white males, rather than seeing how the probability of being male influences the probability of being white. For this type of data, treat it as a single categorical variable with four levels (white males, white females, black males, black females) rather than two categorical variables with each having two levels (black/white, male/female).

Rule #3: Place the treatment/exposure variable as rows and outcome variable as columns.

- Treatment/exposure = causes
- Outcome = effects
  - Not always obvious
  - Example: gestational age

Speaker notes

This relates to the above item. You usually are interested in the probability of an outcome like death or disease, and you are interested in how this probability changes when the treatment or exposure changes. Arranging the table thusly and using row percents usually gets you the comparison you are interested in.

Rule #4: If one variable has a lot more levels than the other variable, place that variable in rows.

"All the News That's Fit to Print"

The New York Times.

LATE CITY EDITION

JAPAN SURRENDERS, END OF WAR! EMPEROR ACCEPTS ALLIED RULE; M'ARTHUR SUPREME COMMANDER; OUR MANPOWER CURBS VOIDED

HIRING MADE LOCAL

Communities, Labor and Management Will Unite Efforts

6,000,000 AFFECTED

Draft Quotas Cut, Services to Drop 5,500,000 in 18 Months

Third Fleet Fells 5 Planes Since End

SECRETS OF RADAR GIVEN TO WORLD

ALL CITY 'LETS GO'

Hundreds of Thousands Roar Joy After Victory Flash Is Received

TIMES SQ. IS JAMMED

Police Estimate Crowd in Area at 2,000,000—Din Overwhelming

PRESIDENT ANNOUNCING SURRENDER OF JAPAN

YIELDING UNQUALIFIED, TRUMAN SAYS

Japan Is Told to Order End of Hostilities, Notify Allied Supreme Commander and Send Emissaries to Him

MACARTHUR TO RECEIVE SURRENDER

Formal Proclamation of V-J Day Awaits Signing of Those Articles—Cease-Fire Order Given to the Allied Forces

Orders Given to the Japanese

PETAIN CONVICTED, SENTENCED TO DIE

Terms Will Reduce Japan To Kingdom Perry Visited

TREATY WITH CHINA SIGNED IN MOSCOW

World News Summarized

Hirohito on Radio; Minister Ends Life

Two-Day Holiday Is Proclaimed; Stores, Banks Close Here Today

Cruiser Sunk, 1,196 Casualties; Took Atom Bomb Cargo to Guam

MacArthur Begins Orders to Hirohito

:::notes A table that is tall and thin is usually easier to read than a table that is short and wide. It is

Rule #5: Whenever you report percentages - always round.

- Two significant figures
  - Exception: close to 100% or 10% or 1%
    - 99.7832% -> 99.8%
    - 10.1417% -> 10.1%
    - 1.2067% -> 1.21%
- Don’t worry about whether your percentages add up to 99% or 101%.

Speaker notes

A change on the order of tenths of a percent are almost never interesting or important. Displaying that tenth of a percent makes it harder to manipulate the numbers to see the big picture.

Adding up to something other than 100% is not a problem. First of all, it can’t happen with a two by two table unless you round incorrectly. For a larger table, it can happen, but your audience is sophisticated enough to understand why this is the case. No one, for example, is going to be upset when 33% plus 33% plus 33% adds up to less than 100%.

Rule #6 When in doubt, write out your table several different ways.

- The fault of default principle
- Revise your tables as often as you rewrite your text

Speaker notes

Pick out the one that gives the clearest picture of what is really happening. Don’t rely on the first draft of your table, just like you would never rely on the first draft of your writing.

A simple fictitious example

Table of counts

	Happy	Miserable	Total
Rich	30	10	40
Poor	90	70	160
Total	120	80	200

Speaker notes

We classify people by their income (rich/poor) and also by their attitude (happy/miserable). There are, for example, 30 rich happy people in our sample and 70 poor miserable people.

Column percents

Table of column percents

	Happy	Miserable	Total
Rich	25%	12%	20%
Poor	75%	88%	80%
Total	100%	100%	100%

Speaker notes

This figure shows column percentages. We compute this by dividing each number by the column total.

We see for example that only 25% of all happy people are rich. This is a conditional probability and is usually written as P[Rich | Happy]. Read the vertical bar as “given.” So this probability is read as the probability of being rich given that you are happy.

Row percents

Table of row percents

	Happy	Miserable	Total
Rich	75%	25%	100%
Poor	56%	44%	100%
Total	60%	40%	100%

Speaker notes

This figure shows row percentages. We compute this by dividing each number by the row total.

We see, for example that 75% of rich people are happy. This is a different conditional probability, P[Happy | Rich]. Read this as the probability of being happy given that you are rich.

Notice the distinction between the two probabilities. Only a few happy people are rich, but most rich people are happy.

Cell percents, 1

Table of cell percents

	Happy	Miserable	Total
Rich	15%	5%	20%
Poor	45%	35%	80%
Total	60%	40%	100%

Speaker notes

This figure shows cell percentages. We compute this by dividing each number by the grand total. Each percentage represents the probability of having two conditions. For example, there is a 15% chance of being rich and happy.

Cell percents, 2

Alternate display of cell percents

Poor and happy	45%
Poor and miserable	35%
Rich and happy	15%
Rich and miserable	5%

Speaker notes

This is an alternate way of displaying cell percentages.

If we had a six categories for attitude rather than just two, we might arrange the table differently.

Handling many categories

Table with many rows

	Rich	Poor
Cloud nine	30% (14)	70% (32)
Cheerful	27% (11)	73% (30)
Content	20% (7)	80% (28)
Despondent	16% (5)	84% (26)
Dejected	11% (3)	89% (24)
Depressed	9% (2)	91% (20)
Total	25% (40)	75% (160)

Speaker notes

Notice that this table would not require any sideways scrolling.

Rules for tables

1. Never display more than one type of number in a table.
2. Row percentages are usually best.
3. Place the treatment/exposure variable as rows and outcome variable as columns.
4. If one variable has a lot more levels than the other variable, place that variable in rows.
5. Whenever you report percentages, always round.
6. When in doubt, write out your table several different ways.

Break #4

- What you have learned
  - Bar plots for categorical outcomes
- What's coming next
  - R code for bar plots

Data dictionary for gardasil, 1

```
---
data_dictionary: gardasil.csv, gardasil.tsv

source:
  This data file is part of the data
  archive for the Journal of Statistics
  Education. The entire archive is at
  http://jse.amstat.org/jse_data_archive.htm


description:
  This data set shows information about young
  women who received the Gardasil shot. Of
  particular interest is the proportion of
  women who received all three shots.
```

Speaker notes

Here is a dataset you will need for your programming assignment. It is a study of pulmonary function in children.

Data dictionary for gardasil, 2

```
additional_description:
  http://www.amstat.org/publications/jse/v19n1/gardasil.txt


download_url:
  http://www.amstat.org/publications/jse/v19n1/gardasil.xls
  http://www.pmean.com/15/images/day2gardasil.csv


copyright:
  Unknown. You should be able to use this data for
  individual educational purposes under the Fair Use
  guidelines of U.S. copyright law.

format:
  csv: comma-delimited
  tsv: tab-delimited

varnames:
 first row of data
```

Speaker notes

Data dictionary for gardasil, 3

```
missing_value_code:
  not needed

size:
  rows: 1413
  columns: 10
```

Speaker notes

t.

Data dictionary for gardasil, 4

```
vars:
  Age:
    label: the patient's age
    unit: years
  AgeGroup:
    label: the age group in which the patient falls
    values:
      11-17 years: 0
      18-26 years: 1
  Race:
    label: the patient's race
    values:
      white: 0
      black: 1
      Hispanic: 2
      other/unknown: 3
```

Speaker notes

t.

Data dictionary for gardasil, 5

```
Shots:
  label: the number of shots that the patients completed

Completed:
  label: did the patient complete the three-shot regimen
  values:
    no: 0
    yes: 1

InsuranceType:
  label: the type of insurance that the patient had
  values:
    medical assistance: 0
    private payer: 1
    hospital based [EHF]: 2
    military: 3

MedAssist:
  label: sid the patient have some type of medical assistance
  values:
    no: 0
    yes: 1
```

Data dictionary for gardasil, 6

```
  Location:
    label: the clinic that the patient attended
    values:
      Odenton: 1
      White Marsh: 2
      Johns Hopkins Outpatient Center: 3
      Bayview: 4

  LocationType:
    label: was the clinic in a suburban or an urban location
    values:
      suburban: 0
      urban: 1

  PracticeType:
    label: the type of practice that the patient visited
    values:
      0: pediatric
      1: family practice
      2: OB-GYN
---
```

simon-5501-04-gardasil.qmd, 1

```
---
title: "Analysis of gardasil shots by demographic factors"
format:
  html:
    embed-resources: true
---

This program reads data on Gardasil vaccinations in young women. Find more
information in the [data dictionary][dd].

[dd]: https://raw.githubusercontent.com/pmean/datasets/master/gardasil.yaml

The program was written by Steve Simon on 2024-09-07 and is placed in the
public domain.
```

Speaker notes

The first few lines are the documentation header.

simon-5501-04-gardasil.qmd, 2

````
## Load the tidyverse library

For most of your programs, you should load the tidyverse library. The messages
and warnings are suppressed.

```{r setup}
#| message: false
#| warning: false
library(tidyverse)
```
````

Speaker notes

h.

simon-5501-04-gardasil.qmd, 3

````
## Read the data and view a brief summary

Use the read_csv function to read the data. The glimpse function will produce
a brief summary. Use tolower to convert uppercase to lowercase.

```{r read}
gard <- read_csv(
  file="../data/gardasil.csv",
  col_names=TRUE,
  col_types="nnnnnnnnnn")
names(gard) <- tolower(names(gard))
glimpse(gard)
```
````

Speaker notes

I.

simon-5501-04-gardasil.qmd, 4

````
## Create factors for agegroup

The factor function identifies a variable as categorical and assigns labels to
number codes. You don't necessarily need to use factor if the data you read in
is character strings, as R automatically treats those variable as categorical.

```{r agegroup-1}
gard$agegroup <- factor(
  gard$agegroup,
  levels=0:1,
  labels=c(
    "11 to 17 years",
    "18 to 26 years"))
```
````

Speaker notes

Use the read_tsv function when your data uses tab delimiters.

simon-5501-04-gardasil.qmd, 5

````
## Counts and percentages for agegroup

```{r agegroup-2}
gard |>
  count(agegroup) |>
  mutate(total=sum(n)) |>
  mutate(pct=round(100*n/total))
```

There are roughly the same number of patients 11 to 17 years as there are
patients 18 to 26 years.
````

Speaker notes

t.

simon-5501-04-gardasil.qmd, 6

````
## Create factors for shots

It is a bit silly to replace 1, 2, 3 with One, Two, Three. The main reason is
to clearly identify shots as categorical rather than continuous.

```{r shots-1}
gard$shots <- factor(
  gard$shots,
  levels=1:3,
  labels=c(
    "One",
    "Two",
    "Three"))
```
````

Speaker notes

t.

simon-5501-04-gardasil.qmd, 7

````
## Counts and percentages for shots

```{r shots-2}
gard |>
  count(shots) |>
  mutate(total=sum(n)) |>
  mutate(pct=round(100*n/total))
```

Slightly more patients got three shots than one or two shots, but this is
still less than half of the patients overall.
````

Speaker notes

t.

simon-5501-04-gardasil.qmd, 8

````
## Compare number of shots by age group

```{r shots-by-age-1}
gard |>
  count(agegroup, shots) |>
  group_by(agegroup) |>
  mutate(row_total=sum(n)) |>
  mutate(pct=round(100*n/row_total))
```
````

Speaker notes

t.

simon-5501-04-gardasil.qmd, 9

````
## Bar chart of shots by age group

```{r shots-by-age-2}
gard |>
  ggplot(aes(x=agegroup, fill=shots)) +
    geom_bar(position="fill") +
    xlab("Age group") +
    ylab("Proportion") +
    ggtitle("Plot drawn by Steve Simon on 2024-09-07")
```

The probability of getting all three shots was higher in the 11 to 17 year old group compared to the 18 to 26 year old group.
````

Speaker notes

t.

Summary

- What you have learned
  - Scatterplots and correlations
  - Boxplots and effect sizes
  - R code for scatterplots and boxplots
  - Bar plots for categorical outcomes
  - R code for bar plots