Cancer data visualization and analysis

Jeff Oliver

21 June, 2021

A two-hour workshop for participants in STEP-UP summer program on cancer prevention and control. Designed for hands-on implementation with a class size of 15-20 students, mostly upper-division undergraduates.

Learning objectives

1. Describe "tidy data" principles of one observation per row, one data type per column
2. Explain how to show relationships in data with data visualization
3. Develop hypotheses to explain quantitative data patterns
4. Write code to visualize data and test hypotheses
5. Explain the difference between correlation & causation

Getting started

- Start RStudio
- Create a new project via File > New Project… > New Directory > New Project
- Make `data` and `output` directories in the project

```
# Create two folders
dir.create("data")
dir.create("output")
```

- Download data (come from https://statecancerprofiles.cancer.gov/)
  - These data are lung cancer data incidence rates for each state, along with some demographic data

```
# Download the data file from the web
download.file(url = "tinyurl.com/cancer-data-csv",
              destfile = "data/cancer-data.csv")
```

Open the file in a spreadsheet program like Microsoft Excel, LibreOffice Sheets, or Google Sheets and look at the data.

Notice that we have 8 columns of data. These are in what is called 'tidy' format. That is because each row has the data for a single state and each column has only one kind of data in it.

Some of these column names are easy to interpret, others are not as useful. The names are very short and it makes them easy to do analyses with them, but they are not necessarily human friendly. So we need to also download the data dictionary that defines what each of those columns means.

```
# Download data dictionary
download.file(url = "tinyurl.com/data-dictionary-csv",
              destfile = "data/data-dictionary.csv")
```

Open this file in your spreadsheet program and see what the columns represent.

column	name	description
state	State	State name
male.lung	Incidence of lung cancer in males	Number of cases per 100,000 males per year
female.lung	Incidence of lung cancer in females	Number of cases per 100,000 females per year
income	Income	Median household income in U.S. dollars
poverty	Poverty	Percent of families below poverty threshold
uninsured	Uninsured	Percent of people 18-64 without health insurance
unemployed	Unemployed	Percent of civilians over 16 years old without a job
lang.isolation	Language isolation	Percent of households where all members have at least some difficulty with English

Now, back in R, we can start working with the data. At this point, we want to make sure we keep track of all the work we do. We can do this by placing all of our R commands in an R script. This is just a fancy way of saying we will type R commands in a text file and save that file for later use. You can create a new script from the file menu (File > New File > R Script). We need to do two more things before we start typing code. The first is to add some information at the top of the script that is for human eyes only. That is, we need to make notes so anyone looking at this file knows what it is supposed to do. So at the very top of your script, starting each line with a pound sign ("#"), add:

1. A short description of what the script does; no longer than one line
2. Your name (not *my* name)
3. Your e-mail address (again, not *my* e-mail address)
4. Today's date in ISO format, YYYY-MM-DD.

```
# Analyze cancer incidence data
# Jeff Oliver
# jcoliver@arizona.edu
# 2020-05-29
```

Second, we need to save our file. Let us save the file under the name "cancer-tests.R".

Now we can load the data into R so we can do our analyses.

```
# Analyze cancer incidence data
# Jeff Oliver
# jcoliver@arizona.edu
# 2020-05-29

# Read data into R
cancer_data <- read.csv(file = "data/cancer-data.csv")
```

We can also take a quick glance at some of these data in r with the `head` and `summary` commands.

```
# Show the first six rows of data
head(cancer_data)
```

```
##        state male.lung female.lung income poverty uninsured unemployed lang.isolation
## 1    Alabama      89.0        51.6  44758    14.0      13.8        8.3            1.2
## 2     Alaska      65.3        50.1  74444     7.0      17.5        7.8            2.2
## 3    Arizona      54.7        45.0  51340    12.9      13.6        8.0            4.5
## 4   Arkansas      98.7        61.6  42336    13.8      11.6        6.9            1.6
## 5 California      49.2        39.0  63783    11.8      10.5        8.7            9.4
## 6   Colorado      46.9        40.7  62520     8.1      10.2        6.0            3.0
```

```r
# Look at the summary statistics for each column
summary(cancer_data)
```

```
##     state              male.lung       female.lung        income         poverty
##  Length:51          Min.   : 32.30   Min.   :23.70   Min.   :40528   Min.   : 5.30
##  Class :character   1st Qu.: 63.17   1st Qu.:50.52   1st Qu.:49037   1st Qu.: 8.10
##  Mode  :character   Median : 69.85   Median :53.45   Median :54384   Median :10.20
##                     Mean   : 72.29   Mean   :53.47   Mean   :56031   Mean   :10.37
##                     3rd Qu.: 82.85   3rd Qu.:58.65   3rd Qu.:62519   3rd Qu.:12.60
##                     Max.   :112.80   Max.   :79.00   Max.   :76067   Max.   :17.40
##                     NA's   :1        NA's   :1
##    uninsured      unemployed    lang.isolation
##  Min.   : 3.7   Min.   :2.800   Min.   :0.300
##  1st Qu.: 7.7   1st Qu.:5.750   1st Qu.:1.500
##  Median :11.2   Median :7.100   Median :2.400
##  Mean   :11.1   Mean   :6.859   Mean   :3.031
##  3rd Qu.:13.7   3rd Qu.:8.050   3rd Qu.:4.200
##  Max.   :22.3   Max.   :9.600   Max.   :9.400
##
```

So what?

But what can we actually do with these data? Well, a lot, really. It really depends on the question you are interested in asking. So one thing we might be interested in is how language isolation affects the rates of lung cancer. That is, do states with a higher percentage of households that have difficulty with the English language have higher rates of lung cancer? This might happen because public health efforts may only be offered in English, and thus not reach all groups in need.

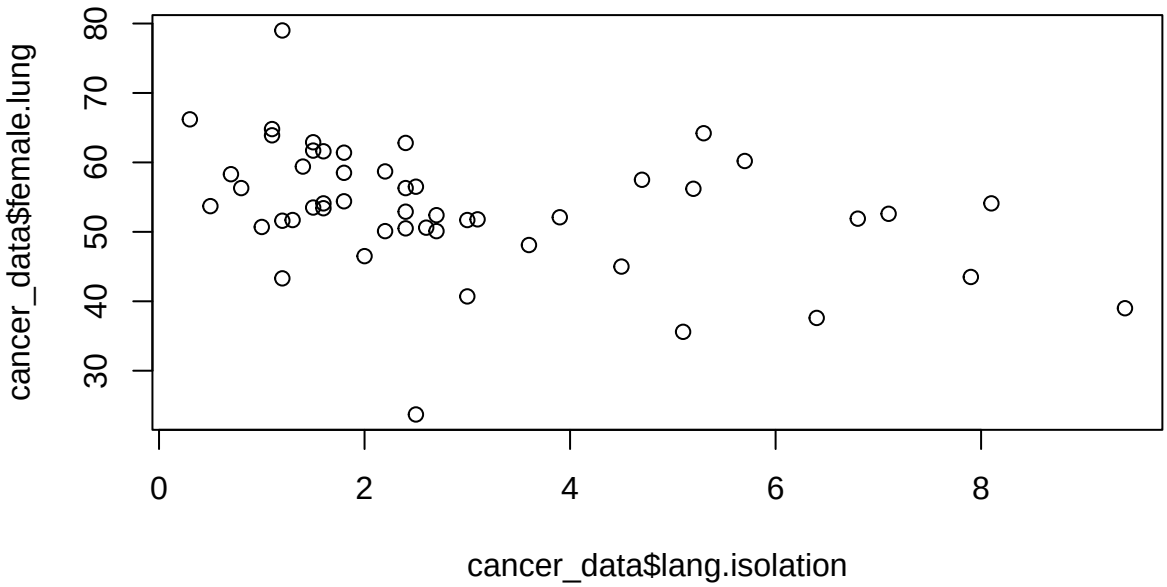
Visualizing data

What would the data look like if this was the case? That is, if we had a plot of language isolation and rates of lung cancer, what would it look like if increased language isolation leads to increased rates of lung cancer? Take a few minutes to draw a plot by hand to show what this looks like.

We can wait.

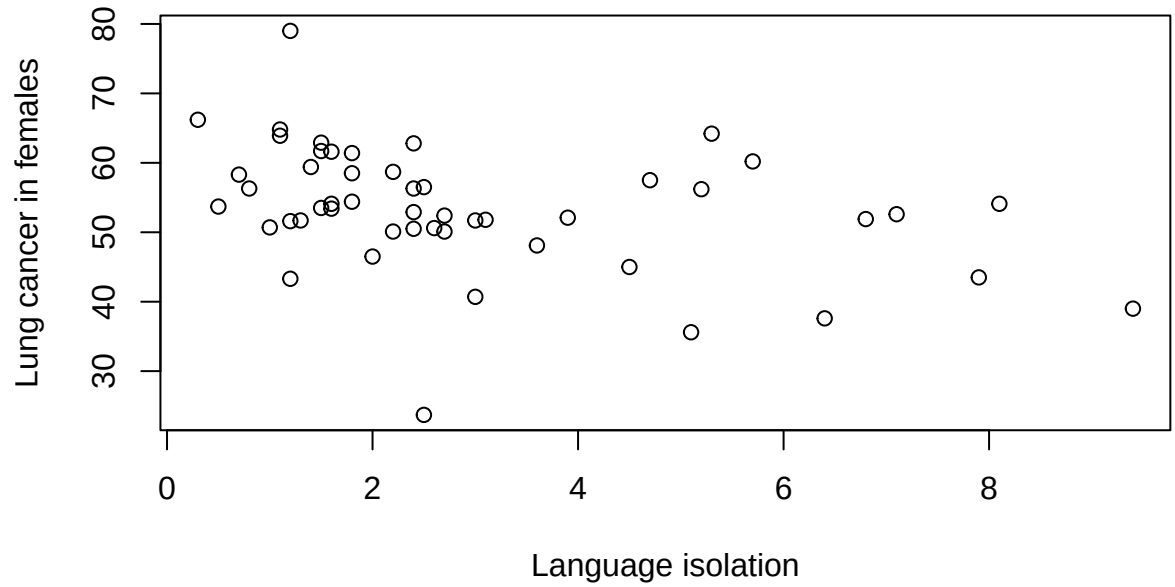
So let us now look at the actual data. For this first plot, we will look at the incidence of lung cancer in females alone. We can use the `plot` command to show the data, instructing R to use the `lang.isolation` column for the x-axis and the `female.lung` column for the y-axis.

```r
# Plot female lung cancer incidence vs language isolation
plot(x = cancer_data$lang.isolation, y = cancer_data$female.lung)
```



Let us take a moment to make our axis labels a little prettier by re-running the plot code with the names we want to use for the x- and y-axes.

```
# Plot again with better axis labels
plot(x = cancer_data$lang.isolation, y = cancer_data$female.lung,
     xlab = "Language isolation", ylab = "Lung cancer in females")
```



Hmmm…it is tough to tell from this plot alone if there is any predictive relationship between language isolation and the incidence of female lung cancer. Regardless of how this plot actually looks, we will want to run a statistical test to see if the relationship is statistically significant.

Generating hypotheses

Before we do that though, let us again take a look at the demographic data we have. We are going to take a moment to generate a hypothesis for one of the other demographic variables. So, for one of the remaining variables (Income, Poverty, Uninsured, Unemployed), consider how it might affect the incidence of lung cancer. If we consider our earlier investigation, the hypothesis we have is:

The rate of lung cancer in females will increase as the amount of language isolation increases.

Exercise Considering how the remaining demographic variables relate to health care disparities, take a moment to write out a hypothesis.

Now that you have your hypothesis, share it with your neighbor.

Testing hypotheses

Let us now return to our original hypothesis, where we posit that language isolation can have an affect on lung cancer incidence. To test this hypothesis, we will use linear regression, which tests the relationship between two continuous-values variables. For the R code, we use the function `lm` to run the analysis and store the output in a variable called `lang_female_lm`.

Aside: When naming variables, we want to make sure the names are actually useful, instead of things like `x` and `var`, which do not say much about what we are storing in them.

```
# Test for an effect of language isolation on female lung cancer incidence
lang_female_lm <- lm(cancer_data$female.lung ~ cancer_data$lang.isolation)
```

We can then look at the output of the linear regression that was stored in `lang_female_lm` by typing the name of the variable alone and running that line:

```
lang_female_lm
```

```
##
## Call:
## lm(formula = cancer_data$female.lung ~ cancer_data$lang.isolation)
##
## Coefficients:
##                (Intercept)  cancer_data$lang.isolation
##                     58.026                      -1.538
```

The output shows the predicted relationship between our two variables, in terms of intercept and slope, but it does not show us whether or not this is a significant relationship. To retrieve that information, we need to use the `summary` function on the results of our linear regression:

```
# View results of linear regression model
summary(lang_female_lm)
```

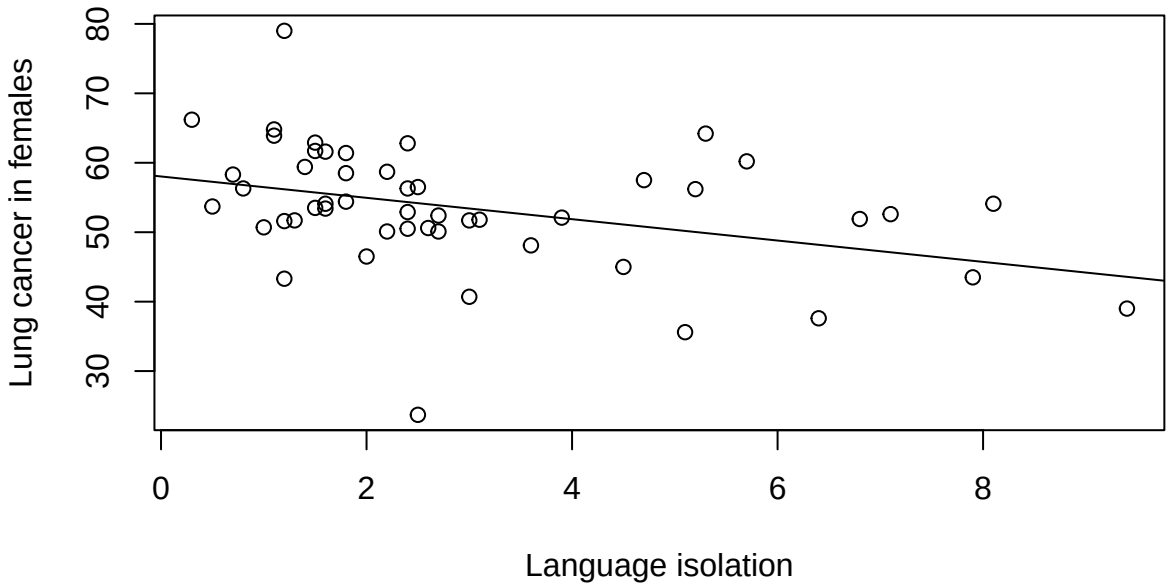
```
##
## Call:
## lm(formula = cancer_data$female.lung ~ cancer_data$lang.isolation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.483  -4.205  -1.147   6.020  22.819
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  58.0264     2.0339  28.529  < 2e-16 ***
## cancer_data$lang.isolation  -1.5376     0.5543  -2.774  0.00786 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.469 on 48 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1382, Adjusted R-squared:  0.1202
```

```
## F-statistic: 7.696 on 1 and 48 DF,  p-value: 0.007862
```

The important part to look at (for our purposes) is the `Coefficients:` section. This shows us the estimated values for the model. In particular, we can see two things about the relationship between language isolation and lung cancer incidence: first, the relationship is *negative*, meaning that lung cancer incidence in females actually *decreases* as language isolation increases. Second, we see this is a statistically significant relationship (because the value of `Pr(>|t|)` is less than 0.05).

We can now use this model to add a line to the plot showing this relationship. We can copy and paste our plotting code from before and add the `abline` function to add the line:

```
# Add the trendline to the plot
plot(x = cancer_data$lang.isolation, y = cancer_data$female.lung,
     xlab = "Language isolation", ylab = "Lung cancer in females")
abline(lang_female_lm)
```



The plot now does a pretty good job of showing the relationship between the incidence of lung cancer in females and the degree of language isolation.

Test *your* hypothesis

Now it is time to test the hypothesis you came up earlier. To do so, you will want to write code that:

1. Runs a statistical test on the data
2. Displays the result of the statistical test
3. Plot the data
4. Add the line from the statistical test to the plot

Which kind of sounds like a lot, but we already did all of this. If we look at our script, we should have something like:

```
# Analyze cancer incidence data
# Jeff Oliver
# jcoliver@arizona.edu
# 2020-05-29

# Read data into memory
cancer_data <- read.csv(file = "data/cancer-data.csv")
```

```
# Test relationship between female lung cancer and language isolation
lang_female_lm <- lm(cancer_data$female.lung ~ cancer_data$lang.isolation)

# Display results of statistical test
summary(lang_female_lm)

# Plot the data and add the results of the statistical test
plot(x = cancer_data$lang.isolation, y = cancer_data$female.lung,
     xlab = "Language isolation", ylab = "Lung cancer in females")
abline(lang_female_lm)
```

We do not need to re-write all the code from scratch, we can copy and paste what we need and update it as necessary (the not-so-well-kept secret of programming is that there is *a lot* of copy-paste that happens).

At this point, try copying all the code, pasting into a new script (with an updated header!), and updating it to test your hypothesis.

Run your code and share your results with your neighbor.

To see this process in action, consider a similar hypothesis to the one above, but this time focus on the incidence of lung cancer in males.

Hypothesis: *The rate of lung cancer in males will increase as the amount of language isolation increases.*

```
# Analyze male lung cancer incidence data
# Jeff Oliver
# jcoliver@arizona.edu
# 2020-05-29

# Read data into memory
cancer_data <- read.csv(file = "data/cancer-data.csv")

# Test relationship between male lung cancer and language isolation
lang_male_lm <- lm(cancer_data$male.lung ~ cancer_data$lang.isolation)

# Display results of statistical test
summary(lang_male_lm)

# Plot the data and add the results of the statistical test
plot(x = cancer_data$lang.isolation, y = cancer_data$male.lung,
     xlab = "Language isolation", ylab = "Lung cancer in males")
abline(lang_male_lm)
```

We can also change which predictor we are looking at. In the next example, instead of looking at how language isolation is related to cancer incidence, we test for a relationship between household income and cancer rates.

Hypothesis: *The rate of lung cancer in females will decrease as average household income increases.*

```
# Analyze lung cancer incidence data and income
# Jeff Oliver
# jcoliver@arizona.edu
# 2020-05-29

# Read data into memory
cancer_data <- read.csv(file = "data/cancer-data.csv")
```

```r
# Test relationship between female lung cancer and income
income_female_lm <- lm(cancer_data$female.lung ~ cancer_data$income)

# Display results of statistical test
summary(income_female_lm)

# Plot the data and add the results of the statistical test
plot(x = cancer_data$income, y = cancer_data$female.lung,
     xlab = "Household income", ylab = "Lung cancer in females")
abline(income_female_lm)
```

Additional resources

- An in-depth explanation of linear regression in R
- A linear regression example that uses the ggplot2 package for nicer-looking plots
- A PDF version of this lesson

Questions? e-mail me at jcoliver@arizona.edu.